INFERRING AND UNDERSTANDING ADAPTATION FROM PATTERNS OF GENETIC DIVERSITY

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

TAYLOR AUSTIN KESSINGER

aus Westminster, California, USA

Tübingen, 2015

Taylor Austin Kessinger: *Inferring and understanding adaptation from patterns of genetic diversity,* Dissertation © 2015

TAG DER MÜNDLICHEN QUALIFIKATION:
08.12.2015

DEKAN:
Prof. Dr. Wolfgang Rosenstiel

BERICHTERSTATTER:
1. Prof. Dr. Daniel Huson
2. Dr. Richard Neher

ERKLÄRUNG

Hiermit erkläre ich, dass ich die Arbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben der Quellen kenntlich gemacht sind.

Tübingen, 2015

Taylor Austin Kessinger

ABSTRACT

The central goal of population genetics is to infer the evolutionary history of a population from observed genetic variation. However, the myriad of evolutionary processes often leave ambiguous signatures, so it can be difficult to reconstruct the evolutionary past. Classical methods typically assume that sequence variation is shaped by neutral processes such as generation to generation sampling variance, i.e., genetic drift.

During the course of my doctoral work, I have undertaken several projects aimed at analyzing populations that are dominated not by neutral processes but by adaptive ones. First, the HIV population within an infected patient experiences strong selection due to immune pressure, and a low recombination rate causes beneficial mutations to sweep concurrently and interfere. To this end, I constructed a realistic model for the evolution of HIV and a method for inferring the selection coefficients of beneficial mutations thereby. Second, many tests of natural selection fail to distinguish between demographic expansion and rapid adaptation. Therefore, I developed a novel method that quantifies the collective effect of many mutations in the genome. The method does not depend on assumptions about demography and can indicate whether genetic draft (i.e., widespread hitchhiking) or genetic drift is the major factor shaping neutral variation. Finally, qualitative differences between rapidly adapting and neutrally evolving asexual populations, such as the statistics of their genealogies, are increasingly well understood, so I contributed to a project that extends coalescence in asexual populations to sexual populations. Properties of sexual populations can therefore be reduced to those of asexual populations, with suitably rescaled parameters.

With this work, I have helped to further a broad and current research program that recognizes the critical role of linked selection, interference, and genetic draft in interpreting patterns of genetic diversity.

ZUSAMMENFASSUNG

Das Hauptziel der Populationsgenetik ist, die evolutionäre Vergangenheit einer Population aus der beobachteten genetischen Variation zu rekonstruieren. Jedoch kann die Vielfalt evolutionärer Prozesse vieldeutige Signaturen hinterlassen und diese Inferenz erschweren. Klassische Methoden setzen normalerweise als gegeben voraus, dass die genetische Variation durch neutrale Prozesse, wie zum Beispiel

Sampling-Varianz von Generation zu Generation (genetische Drift), bestimmt wird.

In meiner Doktorarbeit bschreibe ich verschiedene Projekte, deren Zweck war, Populationen zu analysieren, die nicht durch neutrale sondern adaptive Prozesse dominiert werden. Zum Einen ist die Population des HI-Virus innerhalb eines Patienten stärker natürlicher Selektion durch das Immunsystem ausgesetzt. Seltene Rekombination hat zur Folge, dass mehrere adaptive Mutationen miteinander konkurrieren und sich gegenseitig beeinflussen. Ich habe ein realistisches Model für die Evolution des Virus und eine Methode für die Inferenz von Selektionskoeffizienten der adaptiven Mutationen entwickelt. Zum Andern sind viele Tests auf natürlichen Selektion wenig geeignet, zwischen demographischer Expansion und schneller Adaption zu unterscheiden. Deshalb habe ich eine neue Methode entwickelt, welche den kollektiven Effekt natürlicher Selektion auf viele Mutationen im gesamten Genom quantifiziert. Die Methode ist unabhängig von der demographischen Geschichte und erlaubt zu entscheiden ob genetische "Draft" oder genetische Drift der hauptsächliche Faktor ist, der neutrale Variation dominiert. Darüberhinaus werden qualitative Unterschiede zwischen schnell adaptierenden und neutral evolvierenden asexuellen Populationen zunehmend besser verstanden, zum Beispiel die Statistik ihrer Genealogien. Ich habe an der Entwicklung einer Theorie der Koaleszenz in schnell adaptierenden sexuellen Populationen mitgewirkt. Eigenschaften sexueller Populationen konnten durch geeignetes Reskalieren der Parameter auf das äquivalente Problem in asexuellen Populationen zurückgeführt werden. Mit diesen Projekten habe ich zu dem breiten und aktuellen Forschungsprogramm beigetragen, das die wichtige Bedeutung von Genkopplung, Interferenz, und genetischem "Draft" bei der Interpretation der beobachteten genetischen Diversitätsmuster berücksichtigt.

The fact is that most of us are rather ordinary, having a roughly equal assortment of good and bad genes at those loci where the frequencies of the two are approximately equal. None of us is simultaneously Mozart, Newton, da Vinci, Fisher, Haldane, and Wright.

— Warren J. Ewens

ACKNOWLEDGMENTS

Four years ago, I stepped off of a plane in Stuttgart, bright eyed, bushy tailed, and eager to begin my doctoral work. Science can be stressful; on occasion I felt more like an irritable honey badger than an energetic sciurid. But several people conspired to make this experience very enjoyable for me. I would like to thank some of them here.

First on the list is my undergraduate mentor, Dr. Joanna Masel, a wonderful advisor who kick-started my scientific career. I would not be in Tübingen were it not for her: pursuing my doctorate in a small town in southwest Germany was her suggestion.

At the Max Planck Institute, I'm grateful to my fellow Ph.D. student representatives and many other fellow students. Additionally, it has been my privilege to share the "salt mine" with the members of the Neher research group, an extraordinarily talented and amusing bunch of people. I must single out Fabio Zanini, whose computing expertise and willingness to share it consistently saved me many hours of frustration.

I express my gratitude to the members of my thesis advisory committee for their assistance, including Prof. Dr. Daniel Huson, who has always been an accommodating and helpful advisor. Most of all I thank my main advisor, Dr. Richard Neher. I hope I have picked up on some of the habits that make him an excellent scientist. The one for which I am especially grateful is his emphasis on clear, accurate communication. His tutelage has made me a better researcher, but it has definitely made me *much* better at scientific communication. It has been a privilege to be one of his first Ph.D. students.

My stay in the fatherland has been vastly improved by a number of individuals who have made me feel at home. These include weight room acquaintances, gaming buddies, bandmates, miscellaneous friends, and in particular a few extraordinarily welcoming flatmates who helped me adjust to this foreign environment. If you're reading this, you know who you are.

Finally, I'd like to thank my wife, Nicole Tetu. Amid the toil and turmoil of life as a scientist, she has been the one stabilizing factor. Words cannot express my appreciation of her.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 MOTIVATION

The earliest theories of evolution, such as those of Anaximander, Empedocles, and Aristotle [2, 23], were remarkably sparse on the details of the evolutionary process: the idea that organisms changed over time and had descended from earlier forms (i.e., descent with modification) was much more important than the nuts and bolts of *how* change happens, and they did not conceive of evolution as a process that continued today. The lack of a clear, concrete mechanism arguably inhibited the widespread acceptance of evolution, even when (two millennia later) the work of British and French naturalists rekindled interest in evolution, then known as "transmutation". This changed substantially with the work of Darwin [30], who demonstrated that natural selection might not only cull weak individuals from a population (as Blyth [13] had previously suggested) but also positively aid populations in adapting to new or changing environments. Darwin's thesis did not depend on unquantifiable, cryptic, pre-Galilean notions such as "potentiality" but rather was extrapolated from the observable process of artifical selection as practiced by breeders.

Unfortunately, even Darwin's theory was rather simple. He did not have a satisfactory explanation for how new genetic variation is introduced into a population [31], and he had no concept of the importance of genetic drift. But this setback is passed. A century and a half later, the tapestry of evolutionary theory is woven with many fibers: forces such as mutations, drift, population expansion and contraction, separation of populations, founder effects, epistasis, linkage and hitchhiking, sexual selection, and many others must now be reckoned with.

The central goal of population genetics is to use sequence data (or some other kind of discrete heritable variation, such as binary morphological characters) to infer something about the evolutionary history of a population. The multiplicity of forces involved means that this is not a straightforward task. Patterns of standing variation are seldom unambiguous: it often happens that more than one evolutionary scenario is consistent with the sequence data we observe.

Natural selection remains one of the most important forces in evolution and, consequently, in evolutionary theory and population genetics. In addition to being the major deterministic component of evolution and the only one that can directly lead to adaptation, natural selection is tightly linked to the environment in which an organism

finds itself. After all, the environment is ultimately what imposes the constraints that cause some organisms to perform better than others on average. In this way, selection bridges the gap between population genetics and other areas of biology, such as ecology and functional genomics.

A substantial portion of population genetics has historically been devoted to understanding sequence variation in populations of multicellular eukaryotes. These populations are often somewhat similar to ours, characterized by large genomes with many segregating sites, frequent recombination that breaks up correlations between distant loci, and small population sizes. One major suite of methods comes from neutral theory, according to which most sequence variation is neutral and is shaped by genetic drift, i.e., imperfect sampling of the population from generation to generation [67].

Unfortunately, methods that are perfectly appropriate for understanding these populations are sometimes wrongly applied to populations that in many respects are quite different. For example, many pathogens feature infrequent recombination, pervasive strong selection due to a hostile environment, and very large population sizes. In these populations, selection at linked sites is hugely important in shaping neutral variation. There is even some reason to think adaptive processes rather than neutral ones might be critical in shaping neutral variation throughout the tree of life [76].

In my doctoral research, I have focused on the inference of selection from sequence data. The overarching goal has been to detect selection, especially in rapidly adapting organisms, in a way that robustly respects these populations' unique dynamics.

1.2 SYNOPSIS

Chapter 2 begins by outlining the population genetic theory that forms the groundwork for subsequent chapters. I discuss the behavior of neutrally evolving and rapidly adapting populations, including their coalescent properties, patterns of standing variation, and the fates of individual alleles.

In Chapter 3, I present a novel, simple model for inferring selection coefficients based on time series data from acute infection in HIV. This model is highly parsimonious but takes into account the rapid evolution and fairly low recombination rate of the virus. As a result, it gives selection coefficient estimates that are significantly higher than those of earlier studies.

Next, in Chapter 4, I relate a method for estimating the strength of selection in rapidly adapting organisms. It is based solely on the topology of a genealogy, not the branch lengths thereof. As a result, it is insensitive to demographic change, which sets it apart from many previous methods, and its performance coheres with our intuitions

about the effects of selection on genealogies. It is a specific albeit somewhat noisy tool for assaying the presence and importance of natural selection. I employed this method to infer the strength of selection acting on the HA segment of influenza A subtype H3N2 and found evidence that the segment undergoes very rapid adaptation.

The dynamics, coalescent properties, and patterns of variation in rapidly adapting asexual populations are increasingly well understood. For my third project, which I outline in Chapter 5, I helped extend these principles to sexual populations. The extension relies on a simple, robust scaling argument that considers the length of an effectively asexual block in a sexual genome. In this way, the properties of coalescence and genetic diversity in rapidly adapting sexual populations can be reduced to those of asexual ones.

Finally, in Chapter 6, I offer some comments on major results, the broader picture, and future directions.

STATEMENT ON DATA ANALYSIS AND COMPUTATION

During the course of this work, I have liberally made use of several important Python packages, including:

- `numpy` and `scipy` [96]
- `matplotlib` [63]
- `BioPython` [24]and `Bio.Phylo` [120]

Several others are named and acknowledged in the main text.

2

BACKGROUND

2.1 NEUTRALLY EVOLVING POPULATIONS

The *neutral theory* of molecular evolution has proven one of the most useful and successful research paradigms in population genetics. It was initially formulated by Kimura [67], who suggested that the observed level of divergence in mammal hemoglobin is too large to be explained by natural selection. He posited, as a result of this level of divergence, that most amino acid substitutions and hence mutations are likely to be neutral rather than beneficial. Though he invoked the degeneracy of the genetic code and functional unimportance of many amino acid residues as a justification for this statement, subsequent work, e.g. by Ohno [95], suggested that much of the genome is non-coding and capable of accumulating mutations essentially unabated.

The neutral theory has two major components. One, most observed variation between species or between lineages within a species is neutral: beneficial mutations are rare, and deleterious mutations are typically destined for extinction, but neutral alleles occur frequently enough that they substitute often. Two, the dynamics of most neutral alleles are governed by genetic drift, i.e., imperfect sampling from one generation to the next. I will briefly recapitulate this in terms of the Wright-Fisher model for a haploid population, one of the standard models in population genetics, and explore its implications for the genealogy of a population.

2.1.1 *Wright-Fisher model*

Biology is complicated, so it is almost always necessary to deal with vastly simplified models in order to make useful predictions. As we shall see, the *Wright-Fisher model* allows us to understand a great deal about how populations evolve in certain limits.

We begin with the simplest form of the model, where all individuals are equal in fitness. Consider a population of N non-recombining, haploid individuals, and form the next generation according to the following rule: for each individual, choose a random ancestor in the previous generation with probability $1/N$ (equal probability arises from the assumption of equal fitness). The model can easily be adjusted to incorporate selection (by causing fit parents to be selected with probability proportional to their relative fitness), demography, or other processes. A schematic of this process can be seen in Figure 2.1.

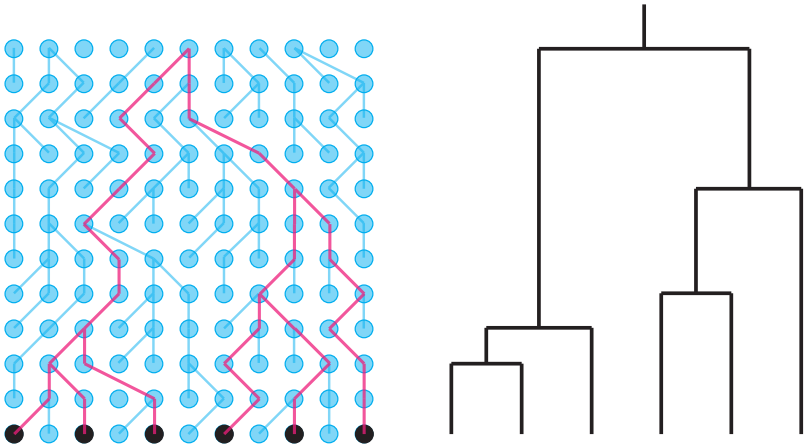

Figure 2.1: A schematic of the Wright-Fisher model (left). Each individual randomly chooses an ancestor in the above generation. We trace the genealogies of different sampled individuals (pink lines of descent). The result is a genealogical tree (right) that is commensurate with the Kingman coalescent.

Herein, I consider two major applications of the neutral Wright-Fisher model, which I will later contrast with a rapidly adapting population: the behavior of neutral variation and the properties of genealogies. I will discuss only the case of asexual haploid individuals. Extensions to diploid individuals are often easily achieved by replacing N with $2N$.

Suppose our haploid population has two variants at some locus, one wild type and one mutant. Let i be the number of mutant individuals at time t (so that the frequency nu is i/N) and let j be the number of mutant individuals at time $t+1$, i.e., the next generation. Then the transition probability matrix relating i and j is given by

$$p_{ij} = \binom{N}{j} \left(\frac{i}{N}\right)^{j} \left(\frac{N-i}{N}\right)^{N-j}. \tag{2.1}$$

This is just a binomial distribution with N trials and "success" probability $i/N = \nu$. Thus, for any value of ν,

$$\mathrm{E}(\Delta\nu) = 0 \tag{2.2}$$

and

$$\mathrm{Var}(\Delta\nu) = \frac{\nu(1-\nu)}{N}. \tag{2.3}$$

In other words, there is no directional change in ν from generation to generation, but ν will experience small fluctuations due to imperfect sampling: the frequency will "drift" with steps on the order of $\sim \sqrt{\nu/N}$. It is worth noting that the variance is *zero* when $\nu = 0$ or 1. These are "absorbing" states: extinct alleles do not re-emerge except via mutations.

Selection is not difficult to incorporate into the Wright-Fisher model. Suppose that, rather than all individuals having equal fitness, the mutant has a selective advantage s over the wild type. The mean offspring number from a mutant individual, i.e., the probability that a mutant individual is selected as the ancestor of a random offspring, will not be 1 but rather $(1+s)/\bar{w}$, with $\bar{w} = \nu(1+s) + (1-\nu) = 1 + s\nu$ the mean fitness. Then

$$p_{ij} = \binom{N}{j} \left(\frac{i(1+s)}{N\bar{w}}\right)^j \left(\frac{N-i}{N\bar{w}}\right)^{N-j}, \tag{2.4}$$

which implies

$$\mathrm{E}(\Delta\nu) = \frac{s\nu(1-\nu)}{\bar{w}} \tag{2.5}$$

and

$$\mathrm{Var}(\Delta\nu) = \frac{\nu(1-\nu)(1+s)}{N\bar{w}^2}. \tag{2.6}$$

Note that, in a continuous limit, we can do away with $\bar{w}$ with no problems. This corresponds to rescaling time so that the mean growth rate in the population is always 1.

At small ν, $1-\nu$ is effectively 1, meaning that ν increases roughly as $s\nu$ per generation: but fluctuations due to drift are on the order of $\sqrt{\nu/N}$. This means that, when ν is below $1/Ns$, drift cannot be neglected in the dynamics. A similar rule applies when ν rises above $1-1/Ns$. In terms of the number of individuals, a beneficial mutation with selection coefficient s needs to drift up to $1/s$ individuals to be assured that it will fix. At this point, the change in the number of mutant individuals $N\nu s$ due to selection is substantially larger than the change due to drift.

The aforementioned general patterns become more obvious when we consider the behavior of alleles in the limit of large population size, i.e., considering the limit as $N \to \infty$ and the time step t becomes small, with Nt kept constant. If $f(\nu, t)$ is the (time dependent) probability density of ν, and if s is small and N large, then the time evolution of $f(\nu, t)$ can be represented as

$$-\frac{\partial}{\partial t} f(\nu,t) = -\frac{1}{N}\frac{\partial^2}{\partial \nu^2}\left[\nu(1-\nu)f(\nu,t)\right] + s\frac{\partial}{\partial \nu}\left[\nu(1-\nu)f(\nu,t)\right]. \tag{2.7}$$

This is known as the diffusion approximation [66]. The $1/N$ prefactor in the first right hand side term sets the scale of diffusive "jumps" due to drift, and the second term corresponds to selection.

One important relation that can be derived from a diffusion approximation is the fixation probability $u(\nu)$ given that an allele with selection coefficient s is currently at frequency ν:

$$u(\nu, s) = \frac{1 - e^{-2Ns\nu}}{1 - e^{-2Ns}}. \tag{2.8}$$

Several limits are worth considering. If $\nu = 1/N$, i.e., the allele is present in one individual, the probability becomes $(1 - e^{-2s})/(1 - e^{-2Ns})$. In this limit, if s is small but Ns is large, the bottom exponent $\to 0$ and the top $\to 2s$. (In some alternate models such as the branching process model of Desai and Fisher [32], the fixation probability in this limit is s rather than $2s$.) If s is zero, the probability becomes $1/N$ (after applying L'Hôpital's rule). If s is negative, the fixation probability quickly asymptotes toward zero. Thus, beneficial alleles stand a good chance of fixing, whereas deleterious ones are very unlikely to do so.

As one might expect, not every beneficial mutation is destined to fix, i.e., reach frequency 1. Rather, beneficial mutations below $1/Ns$ experience substantial fluctuations, meaning there is a high probability that they will be lost to drift. Provided that one is ultimately destined to fix, it persists in the population for roughly $1/s$ individuals before passing the drift barrier.

Suppose a beneficial mutation has reached $\nu \approx 1/Ns$, so that its dynamics begin to be governed primarily by selection rather than drift. This condition is known as *establishment*. Then stochastic effects can be ignored, and the expression for $\mathrm{E}(\Delta\nu)$ above yields a differential equation,

$$\frac{d}{dt}\nu = s\nu(1-\nu), \tag{2.9}$$

which has the solution

$$\nu(t) = \frac{\nu_0 e^{st}}{1 + \nu_0(e^{st} - 1)}, \tag{2.10}$$

with $\nu_0 = 1/Ns$. The time for the mutation to sweep through the bulk of the population, i.e., to transition from $\nu = 1/Ns$ to the other drift barrier at $1 - 1/Ns$, is obtained by solving the above equation for $\nu = 1 - 1/Ns$. This yields

$$t = \frac{2}{s}\log(1 - \frac{1}{Ns}) + \frac{2}{s}\log(Ns). \tag{2.11}$$

If the product Ns is small, the former term can be ignored, so that the sweep time scales simply as $\sim s^{-1}\log(Ns)$. This is often a very short time interval. If the rate at which beneficial mutations enter the population is U, then the total number entering the population every generation is NU: so in general, a beneficial mutation will fix every $1/NUs$ generations. Provided this is shorter than the typical sweep time $s^{-1}\log(Ns)$, *beneficial mutations do not interfere*: a beneficial mutation is fixed by the time the next one arrives in the population. Each beneficial mutation increases the mean fitness by s as it sweeps, so fitness increases at a total rate $v \sim NUs^2$.

As an extremely general rule, genetic drift decreases genetic diversity: it can cause loci to hit the absorbing barriers at 0 or 1, but it

cannot reverse this process. Consider p, the probability that two randomly sampled gene copies at a locus are different (this is sometimes referred to as the "heterozygosity"). Then the expected value of p, when ν is the frequency of a mutant allele, is simply

$$p = 2\nu(1-\nu). \tag{2.12}$$

What happens to p as an allele's frequency changes due to genetic drift? Let p' and ν' be the values of p and ν in the next generation. Then

$$\mathrm{E}(p') = \mathrm{E}(2\nu'(1-\nu') = 2\nu(1-\nu)(1-\frac{1}{N}) = p(1-\frac{1}{N}). \tag{2.13}$$

In general, t generations into the future, we have

$$\mathrm{E}(p_t) = p\left(1-\frac{1}{N}\right)^t \approx pe^{-\frac{t}{N}}. \tag{2.14}$$

That is to say, genetic drift causes a decay in variation, and at long time scales or in small populations, it is all but certain that even a neutral allele will disappear from the population.

2.1.2 *Kingman coalescence*

The Wright-Fisher model is not only useful for understanding the forward-time dynamics of a population. With a little work, we can rewind the tape and model the genealogical history of a population, as well.

Consider b individuals sampled from a population of size N. What is the probability that, in the previous generation, b individuals have exactly b distinct ancestors? Essentially one needs to throw b balls into b distinct boxes out of N. When the first ball is thrown, the second will land in a different box with probability $1-1/N$: the third will land in yet another box with probability $1-2/N$: and so on. Overall, the probability becomes

$$\begin{aligned} P(b \text{ distinct boxes}) &= \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\left(1-\frac{b-1}{N}\right) \\ &= 1-\sum_{j=1}^{b-1}\frac{j}{N}+O\left(\frac{1}{N^2}\right) \\ &= 1-\frac{\binom{b}{2}}{N}+O\left(\frac{1}{N^2}\right), \end{aligned} \tag{2.15}$$

by tracking only the terms of order $1/N$ or higher. Similarly, two balls land in the same box with probability $1/N$, and there are $\binom{b}{2}$ possible

pairs of balls: so the probability of landing b balls into exactly $b-1$ boxes becomes

$$\begin{aligned} P(b-1 \text{ distinct boxes}) &= \frac{\binom{b}{2}}{N}\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\left(1-\frac{b-2}{N}\right) \\ &= \frac{\binom{i}{2}}{N}+O\left(\frac{1}{N^2}\right). \end{aligned} \tag{2.16}$$

The remaining terms are of order $1/N^2$ and can be ignored. In this way, for large N, the process of individuals coalescing backwards in time is like a series of Bernoulli trials with success probability $b(b-1)/2N$. Taking the limit $N \to \infty$ and considering a very large number of trials (i.e., very small time steps) yields a process where individuals merge at an exponentially distributed rate $b(b-1)/2N$.

Thus, in the limit of large population size, the genealogies of neutrally evolving Wright-Fisher populations can be simulated in the following way: when there are b individuals, wait for a period of time distributed as $\exp(2N/b(b-1))$ and merge two at random. b begins at n, and the process continues until $b=1$, i.e., all lineages have been merged. This is the *Kingman coalescent* [69]: refer to Figure 2.1 for an illustration.

The Kingman coalescent is a powerful tool for modeling the evolution of neutral populations. Individuals are treated as completely exchangeable, i.e., the merging process is identical no matter which labeled individuals belong to a lineage (or even whether the individuals are labeled at all), and coalescence times are independent. It is, additionally, easy to simulate, as only the n sampled individuals need to be considered rather than potentially N individuals at many time points (as in the forward Wright-Fisher model).

The Kingman coalescent provides a framework for analyzing a population's history that can be much simpler to work with than the forward Wright-Fisher model. For example, what is the mean time to coalescence of an entire sample of n individuals? Consider that, when b lineages are extant, the mean wait time until the next branching event is simply $2N/b(b-1)$: thus,

$$\mathrm{E}(T_{\mathrm{MRCA}}) = \sum_{b=2}^{n}\frac{2N}{b(b-1)} = 2N\sum_{b=2}^{n}\left(\frac{1}{b-1}-\frac{1}{b}\right) = 2N\left(1-\frac{1}{n}\right). \tag{2.17}$$

Inserting $n=2$ gives the average pair coalescence time, N. The variance in each coalescence time is simply $[2N/b(b-1)]^2$, and a similar sum yields

$$\mathrm{Var}(T_{\mathrm{MRCA}}) = 4N^2\left[2\sum_{b=2}^{n}\frac{1}{b^2}-\left(1-\frac{1}{n}\right)^2\right]. \tag{2.18}$$

It is worth noting that the two moments have similar behavior. The total length of the tree can likewise be calculated by noting that, when

there are b branches on the tree, the wait time until the next coalescence event $2N/b(b-1)$ multiplied by b gives the branch length in that interval: thus,

$$\mathrm{E}(T_{\text{total}}) = \sum_{b=2}^{n} b\frac{2N}{b(b-1)} = 2N\sum_{b=2}^{n-1}\frac{1}{b} \sim N\log n. \tag{2.19}$$

The variance likewise becomes

$$\mathrm{Var}(T_{\text{total}}) = 4N^2\sum_{b=2}^{n-1}\frac{1}{b^2}. \tag{2.20}$$

2.2 RAPIDLY ADAPTING POPULATIONS

In natural populations, individuals are essentially never equal in fitness. Rather, some mutations convey a fitness advantage and others a disadvantage. The situation that is often considered in population genetics is one where fitness variation is determined by a fairly small number of mutations with large fitness effects, which "sweep" out neutral diversity as they expand in the population. In sexual populations, these loci can be treated as outliers against a neutral genetic background: the selective advantage s determines roughly how quickly the sweep happens, but recombination occurs at rate ρ, decorrelating distant loci from the effects of the beneficial mutant.

It is increasingly realized, however, that this picture may not be accurate. Fitness variation may be due to the effects of multiple competing genetic backgrounds or to "soft sweeps" [49, 59], where different beneficial alleles or copies thereof persist on distinct genetic backgrounds. Selection can be frequent and ubiquitous, and the distinction between sweep and background often does not hold.

Prior work on rapidly adapting populations has typically focused on regimes where drift is occasionally interrupted by strong selective sweeps [6, 123] or strong purifying selection at a small number of sites [122]. Here, we consider the alternate scenario, one where differences in fitness between individuals are due to many loci of fairly small effect so that the fitness distribution is somewhat smooth: in quantitative genetics, this limit is sometimes referred to as the infinitesimal model [21, 82]. Such populations are not only quantatively but also qualitatively different from neutrally evolving populations. We focus on the behavior of neutral alleles and the properties of genealogies: later, I will comment on implications for the effects of population size and patterns of neutral variation.

2.2.1 *Traveling fitness waves*

If fitness variation is due to the additive effects of many loci with small contributions, then by the central limit theorem, the bulk of the

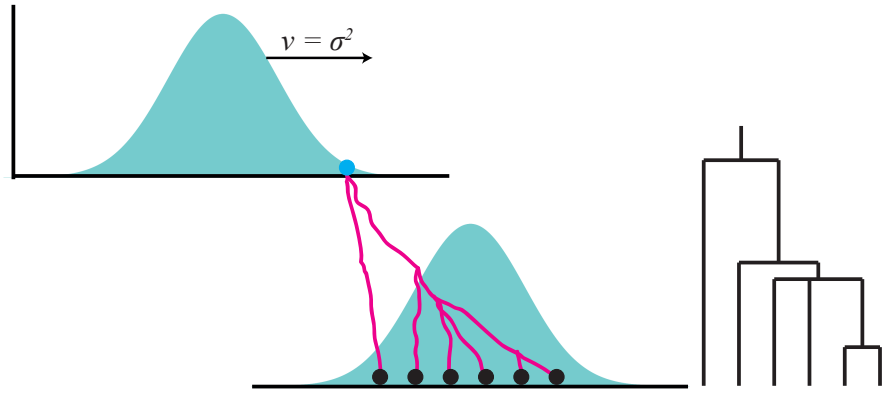

Figure 2.2: The genealogy of a rapidly adapting population. When the fitness distribution is due to many small effect loci, it is approximately Gaussian. The population adapts at a speed $v = \sigma^2$. Individuals quickly trace their lineage back to fit ancestors near the nose of the fitness distribution (left). Such genealogies are well described by the Bolthausen-Sznitman coalescent (right), with multiple mergers and long terminal branches.

fitness distribution can be approximated as a Gaussian. The distribution will not remain stationary, however. Fit individuals near the nose will expand quickly; their ancestors will later comprise the bulk of the population, and unfit individuals in the tail will quickly be driven to extinction. The width of this distribution determines how much fitter the individuals in the nose are than the remainder of the population. As they expand, the bulk of the distribution appears to move forward, and if mutations provide an influx of fitness variation, the bulk shifts ahead as individuals in the nose continue to dominate.

In such a scenario, the fate of neutral alleles is not governed primarily by genetic drift. Rather, it matters on which background a new mutation appears. Mutations that appear in fit individuals, i.e., near the nose, will ride the wave as individuals carrying them give rise to a large chunk of the population. Mutations that appear elsewhere in the population are doomed to extinction.

We can now describe this process with just a bit more rigor. Consider an asexual haploid population of size N with additive fitness variance σ^2. Fisher's "fundamental theorem" of natural selection [44] states that the population's mean fitness $\bar{x}$ will change as

$$\frac{d}{dt}\bar{x} = \sigma^2 \tag{2.21}$$

due to natural selection: see the left portion of Figure 2.2. A more complete expression can be written incorporating the effects of epistatic variance, contributions from individual mutations, and so on: we ignore these for the sake of simplicity. We proceed by considering only the case where the variance is roughly constant and an influx of mutations maintains it.

In this limit, beneficial mutations do not sweep as they might in a neutrally evolving population: rather, they interfere. Recall that when beneficial mutations are reasonably rare, so that a mutation's sweep time $s^{-1} \log Ns$ is much smaller than the rate at which new mutations establish NUs, beneficial mutations spread through the population unhindered. However, that is explicitly not the case in the rapidly adapting populations under consideration. Instead, multiple beneficial mutants may appear on the same background, aiding each other's expansion, or they may appear on different backgrounds, in which case they will interfere. The latter is more likely, meaning that the rate of adaptation is overall slowed: it is logarithmic rather than linear in N and U [32, 50]. Note that frequent recombination can restore the "successive beneficial mutations" dynamics: recombination decorrelates the behavior of sweeping alleles, so double mutants can be produced fairly easily via recombination, and often they can sweep without interference. Interference between beneficial mutants, especially at nearby loci, due to weak recombination is known as Hill-Robertson interference [6, 61] and can be seen as a weaker form of clonal interference.

Like a beneficial mutation, the frequency of a *neutral* mutation is strongly affected not only by sampling variance from generation to generation but also by hitchhiking, i.e., the relative fitness of the genetic background. This widespread hitchhiking is known as *genetic draft* [52]. Loci that appear in fit individuals stand a good chance of fixation. Loci that do not are very unlikely to fix. One consequence of this dependence on genetic background is that the overall fixation probabilities *averaged over the entire population* remain the same, but the fixation probability depends strongly on the fitness x of the genetic background on which it appears. In general, it is exponential in x as one nears the nose of the distribution, then linearly increases in x thereafter [92]. This stands in stark contrast to the situation with genetic drift where the fixation probability for a new neutral mutation is $1/N$ no matter whose genome it appears in.

An important property of rapidly adapting populations in traveling wave and other models is the distribution of the number of offspring. In a neutral Wright-Fisher population it is binomial (and is often approximated as Poisson), with mean and variance 1. A rapidly adapting population can be described by a Wright-Fisher model, but there is a feature that becomes obvious over longer time scales. Consider an asexual population and an individual with fitness x whose lineage is large enough to have established in the population. Then its number of descendants at time t is given roughly by

$$n(x,t) \sim \frac{1}{x} e^{xt - \sigma^2 t^2/2}, \tag{2.22}$$

letting the mean fitness $\bar{x} = 0$ at $t = 0$. The second term in the exponent corresponds to the increasing mean fitness in the rest of

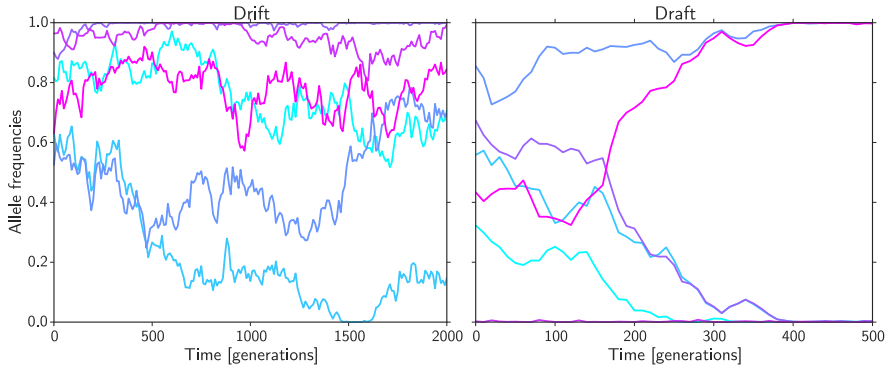

Figure 2.3: The behavior of neutral alleles under genetic drift (left) and genetic draft (right). Under drift, neutral alleles take small, diffusive steps in frequency (of order $1/\sqrt{N}$), and they may persist for a very long time in the population before fixing or going extinct. Under draft, neutral alleles are rapidly swept about by the effects of the underlying genetic background (note the different time scales), and they take large, sustained jumps that do not lend themselves to a diffusion approximation.

the population: $\sigma^2 t^2/2 = \int_0^t \bar{x}(t')dt'$. Since we know the distribution $P(x,0) = (2\pi\sigma^2)^{-1/2}\exp(x^2/\sigma^2)$, we can calculate the distribution $P(n,t)$ by using the above approximation and computing $P(n,t) = P(x,t)dx/dn$ [91]. Doing so yields

$$P(n,t) \sim \frac{1}{n^{3/2}} e^{-\frac{\sigma^2 t^2}{8} - \frac{\log^2 n}{2\sigma^2 t^2}}. \tag{2.23}$$

Integrating over time gives the distribution of the total number of offspring a clone produces, which scales as

$$P(n) \sim \frac{1}{n^2}, \tag{2.24}$$

in stark contrast to the binomial (or approximately Poisson) one-generation offspring number distribution under the Wright-Fisher model. One astonishing feature should now be obvious. The offspring number variance is given by

$$\mathrm{Var}(n) = \int_0^\infty P(n)n^2 dn - \left(\int_0^\infty P(n)n dn\right)^2, \tag{2.25}$$

but both integrals diverge (the first one more quickly), meaning that the offspring number variance likewise diverges [91]. In practical terms, this divergence is prevented by the fact that the population size is finite, meaning that the Gaussianity of the fitness distribution is cut off by the discretization of the number of individuals: the fittest individual is present not at infinite fitness but rather roughly $x_c = \sigma\sqrt{2\log N\sigma}$ [107, 121], which cuts off the integral. However, the distribution is at least fat tailed, meaning that *no diffusion approximation is possible*.

Over shorter time scales, the jumps in the allele frequency ν are likewise fat tailed. Allele frequencies take large jumps due to the underlying ebb and flow of the genetic background that cannot be captured by a diffusion approximation. This key difference between drift and draft is illustrated in Figure 2.3. Over long time scales in a sexual population, it is even possible to construct a modified Wright-Fisher model where the relevant timescale is the average lifespan of clones [91]: here, too, the distribution in the number of recombinant offspring from one clone has a diverging variance.

2.2.2 *The Bolthausen-Sznitman coalescent*

An additional critical difference between rapidly adapting and neutrally evolving populations concerns the shapes of their genealogies. In a rapidly adapting population, fit individuals give rise to disproportionately large chunks of the future population. This means that the Kingman coalescent is not an appropriate model for genealogies, as individuals are no longer fully exchangeable: some organisms are more equal than others. We consider here an alternative coalescent model, the *Bolthausen-Sznitman coalescent* or BSC [14]. It is not straightforward to derive the BSC from a realistic model of adaptation, so we outline its properties first and later argue how it can arise from rapid adaptation. Note that the BSC has been explicitly shown to arise from specific traveling wave models, in particular the "exponential model" of Brunet et al. [20], as well as from populations where some lineages undergo transient rapid expansion [113].

In the Kingman coalescent, when b lineages are present, two of them merge at rate $\lambda_b = b(b-1)/2N$. The Bolthausen-Sznitman coalescent is similar in some respects to the Kingman coalescent, except that it is possible that more than two lineages will merge. In general, k out of b lineages merge at rate

$$\lambda_{b,k} = \frac{(k-2)!(b-k)!}{(b-1)!}. \tag{2.26}$$

Note that this applies to *any* set of k out of b lineages, so that the rate at which any merger of size k happens is $\binom{b}{k}\lambda_{b,k}$, and the total merger rate is

$$\lambda_b = \sum_{k=2}^{b} \binom{b}{k} \lambda_{b,k} = b - 1, \tag{2.27}$$

again in contrast to the Kingman coalescent, where $\lambda_b \sim b(b-1)$. It is easy to see that $\mathrm{E}(k) \approx \log b$ for an arbitrary merger event, i.e., each merger decreases the number of lineages by about $\log b$. Furthermore, the distribution of coalescence times differs: the time for n individuals to coalesce scales not as $1 - 1/n$ as in the Kingman coalescent but as $\log\log n$, i.e., coalescence is much more rapid and less dependent on the number of leaves.

Two results here are important. First, multiple mergers lead to very skewed branchings. In the Kingman coalescent, branchings tend to be very balanced, but the BSC's tendency to merge large swathes of the population in a single step means that some lineages will have many more descendants than others. Second, the merger rate's lessened dependence on leaf number means that the BSC features disproportionately longer terminal branches, including (in some cases) terminal branches that persist deep into the tree.

In effect, differential fitness in the population means that individuals are strictly speaking not exchangeable. The population has a "memory": parent and offspring fitnesses are correlated, so that if one individual gives rise to many offspring, its offspring likely will, too. However, the BSC, like all coalescent models, assumes no correlation between parent and offspring fitnesses, i.e., it treats individuals as exchangeable. This apparent paradox is resolved by the fact that, over fairly long time scales, the fitness of ancestors and descendants does indeed decorrelate, which allows the BSC to be a good approximation.

An intuitive argument can make clear how the BSC, or something like it, must emerge from models of rapid adaptation. As a population moves toward higher fitnesses, individuals in the nose grow and overtake the bulk of the population, essentially becoming the new bulk. This means that extant sampled individuals in the present must coalesce quickly to individuals in the fairly recent past. As very fit individuals expand exponentially, comprising a large portion of the bulk, likewise extant individuals will sometimes coalesce back to these fit ancestors over short time scales. Lastly, unfit individuals can still persist in the population albeit in small number, so their lineages may merge with the bulk deep in the tree. Refer back to Figure 2.2.

The BSC is more typically associated with organisms with very skewed offspring number distributions, in which "sweepstakes reproduction" or frequent (re-)colonization of environments are the norm [37]. However, it emerges from at least one explicit model of adaptation as previously mentioned, and it has been shown to characterize more realistic models of rapid adaptation, both approximately [33, 89] and rigorously [114]. The approximation emerges because true "multiple mergers" are rare in typical adaptation models. Rather, fit individuals expand and come to dominate large chunks of the population over $\sim \sigma^{-1}$ generations.

2.3 THE ROLE OF POPULATION SIZE

In the case of the Kingman coalescent, the Wright-Fisher diffusion, and a variety of relations that can be derived from either one (such as the sojourn time and fixation probability for various mutations or the per-generation loss in variation), the population size plays a

central role. In forward time, the population size sets the scale of diffusive "jumps" in allele frequency, which primarily affects neutral or nearly neutral alleles. In reverse time, the population size sets the rate at which lineages coalesce (and ultimately the entire population coalesces).

However, there are a variety of processes that can slightly alter the dynamics. For example, in a sexual population, mate choice can cause the number of reproducing males or females to be limited. This means the effective breeding number of individuals is less than N. Pervasive background selection or selective sweeps can likewise depress the number of participating individuals. One can also consider the effects of a changing population size. Suppose that N changes at each time step, with values $N_1, N_2, \ldots N_t$. Then in general,

$$\mathrm{E}(p_t) = p \prod_{i=1}^{t} \left(1 - \frac{1}{N_i}\right). \tag{2.28}$$

What is the size of an effective population N_e that loses variation at the same as this one does? Observe that, when N is large, $1 - \frac{1}{N} \approx \exp(-1/N)$: so

$$e^{-t/N_e} = \prod_{i=1}^{t} e^{-1/N_i}, \tag{2.29}$$

or

$$\frac{1}{N_e} = \frac{1}{t} \sum_{i=1}^{t} \frac{1}{N_i}. \tag{2.30}$$

That is, the effective population size is the harmonic mean of the per-generation population sizes. When some process other than demographic change is responsible for depressing the number of breeding individuals, the relationship between N and N_e can be more complicated, but the result is purportedly the same: the population simply evolves as though its size were N_e rather than N.

In this way, a reduced neutral model still describes the population, albeit with a larger diffusion step size and a shorter coalescence time. In most other respects, populations are not treated any differently. It is still assumed, for example, that neutral alleles behave in a well-defined diffusive manner, that similar ratios between pairwise and total coalescence times obtain, and that neutral site frequency spectra decrease monotonically. In general N_e should scale with N, and therefore other statistics such as heterozygosity and the number of segregating sites should likewise scale with N [66, 69].

In natural populations, however, the correlation between N and inferred N_e is surprisingly weak [53], which has been referred to as a "paradox of variation" [76]. One explanation that has been offered for this is that neutral diversity is governed primarily not by genetic *drift* but by *draft* [29, 73]. If true, this accounts very well for the lack of correlation between N and N_e. Purportedly, even though the population's size is really N, it behaves as though its population size is N_e:

genetic drift affects it as though its population size is smaller, with the size of diffusive "jumps" in neutral allele frequencies thereby being larger. But again, when genetic draft is more important than genetic drift, there is no diffusion limit [91], meaning that *there is no effective population size*. Refer back to Figure 2.3. N_e becomes a fudge factor for rationalizing away problems with the standard neutral model, when it is quite possible that an alternative null model is needed.

Likewise, rescaling N to N_e affects the typical coalescence time of populations under the Kingman coalescent. But if coalescence occurs according to an alternative process such as the BSC, the dependence on N is much weaker. For example, in the BSC, the population size dependence of the average pair coalescence time $\langle T_2 \rangle$, which determines the coalescence time scale for the remainder of the tree, turns out to be very weak, proportional to $\sqrt[3]{\log N}$ [89], and the dependence of higher coalescence times on $\langle T_2 \rangle$ is not as straightforward [20]. More importantly, BSC genealogies are different not only in terms of branch length but also in terms of topology, i.e., the pattern of branching events itself: lineages undergo multiple mergers and sometimes persist very deep into the tree unabated. Simply rescaling coalescence times according to N_e cannot capture this feature of a genealogy [111].

The relevant parameter for determining whether a population's genealogy behaves in a Kingman or BSC-like manner is the product of the population size N and the standard deviation in fitness σ [54, 93]. Large values of N ensure that, even if σ is small, the fittest individuals in the population (who will ultimately give rise to much of it) are found far ahead of the bulk: and large values of σ, of course, mean that the population adapts rapidly (via $v = \sigma^2$. The intermediate regime $N\sigma \approx 1$ is relatively unexplored.

As a result of considerations such as these, the battle cry *death to "effective population size!"* can occasionally be heard among population geneticists at conferences. There is clearly a pressing need to develop a novel null model for understanding evolution that does not depend primarily on rescaling the population size.

2.4 SITE FREQUENCY SPECTRA

Population geneticists often seek informative summary statistics, whose values or distributions can be tested against an appropriate null model. The *site frequency spectrum*, a graph that features, on the x axis, the frequency ν of an arbitrary (presumably neutral) derived allele, and on the y axis, the proportion of mutations $f(\nu)$ that are present at that frequency, is one way to summarize the diversity in a population: it incorporates information about useful statistics such as the mean pairwise difference or number of singletons. The general shape of the SFS depends on (and, accordingly, is informative about) the underlying processes governing neutral variation. In fact, many classical tests

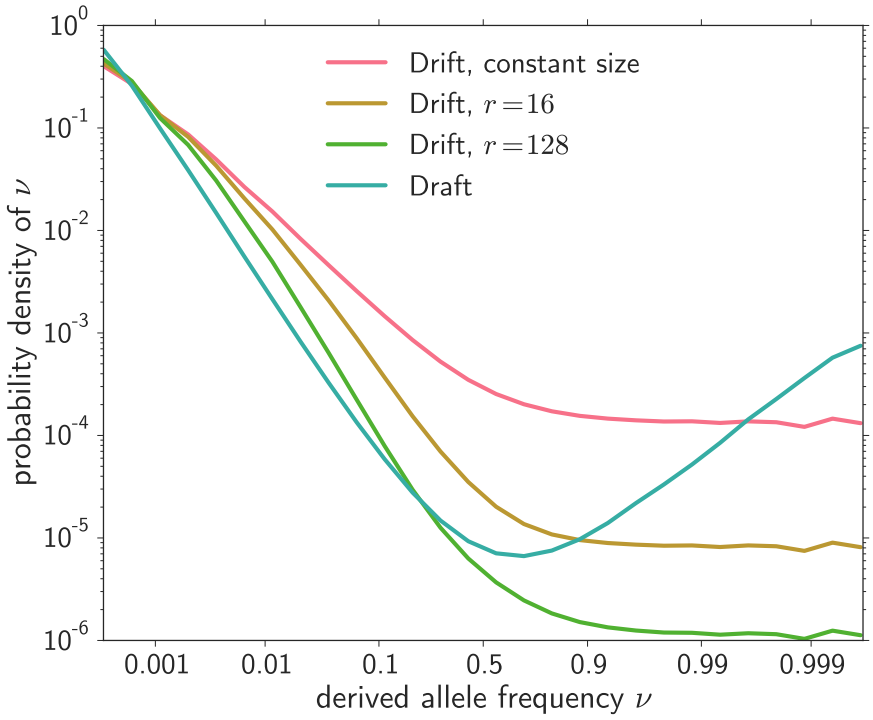

Figure 2.4: The site frequency spectrum for various populations. Drift causes the distribution to descend quickly as ν^{-1}. Population expansion can lead to a steeper spectrum by preventing intermediate frequency or common alleles from accumulating. But the non-monotonicity of the spectrum (excess of common alleles) under draft is something that cannot be duplicated by simple population expansion.

of natural selection are effectively tests of the shape of an observed spectrum.

Under genetic drift, most mutations are doomed to extinction: they appear in the population and persist for a few generations before disappearing. A few lucky mutations may occasionally drift to high frequency, but this is not common. However, if a mutation manages to survive to intermediate frequencies, it is likely to persist for some time. The spectrum scales as $f(\nu) \sim \nu^{-1}$. Demographic effects can distort the spectrum: for example, an expanding population yields a steeper spectrum, as mutations that arose early in the population's history, when the population was small, stood a better chance of drifting up to intermediate frequencies.

Unsurprisingly, genetic draft results in a qualitatively different spectrum. Mutations that arise on backgrounds not destined to dominate the population are themselves destined for extinction. They are swept out quickly, so that at low frequencies, the spectrum scales as $f(\nu) \sim \nu^{-2}$ [18]. However, mutations on fit genetic backgrounds may ride the wave to very high frequencies, so that close to 1, the spectrum

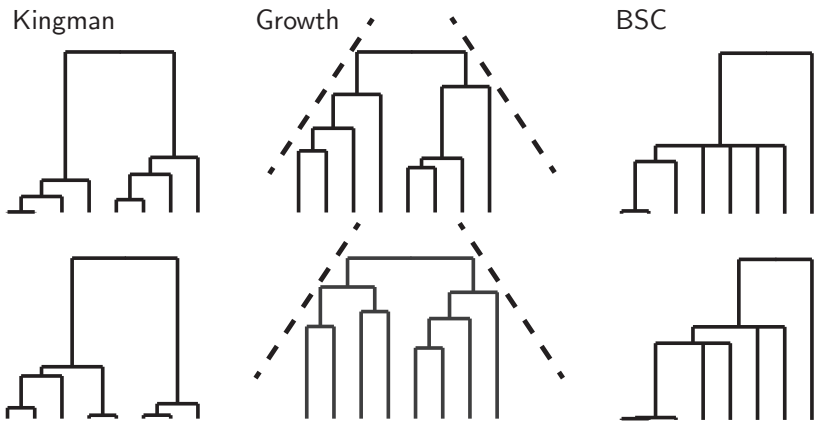

Figure 2.5: Coalescent trees. The Kingman coalescent gives rise to balanced trees in general: the BSC features approximate multiple mergers and skewed branchings. The middle case, representing population expansion, features higher merger rates in the past and hence longer terminal branches. In this way, it can be mistaken for a genealogy associated with rapid adaptation, even though its *topology* does not necessarily differ from the Kingman expectation.

behaves as $f(\nu) \sim 1/(\nu - 1)\log(1 - \nu)$ [89], i.e., the spectrum is non-monotonic and experiences a significant uptick close to $\nu = 1$. An excess of common derives alleles is, therefore, a hallmark of selection. This non-monotonicity can be understood in terms of a depletion of intermediate frequency alleles, as well. There is no reason for an allele in the middle of the spectrum to "stick around". To simplify, either it is on an unfit background, in which case it will disappear shortly, or it is on a fit background, in which case it will fix shortly. Refer back to Figure 2.3.

The shape of the SFS is closely linked to the shape of genealogies, which present a convenient summary of the genetic variation in the population: see Figure 2.5. Novel alleles arise along the branches of the genealogy and are inherited by all downstream individuals. In the Kingman coalescent, branchings tend to be balanced, so there are overall some intermediate frequency alleles, and the total branch length on the tree slowly increases as one moves toward the present, so rare alleles should dominate the spectrum. In the BSC, a substantial portion of the tree's branch length is terminal or near terminal, so rare alleles are present in excess, but branchings tend to be very uneven, meaning that common alleles (inherited by all individuals downstream of the larger branch) are also abundant [71, 89].

The non-monotonicity of the site frequency spectrum under genetic draft is a powerful but underappreciated signature of selection. As I will discuss further in Chapter 4, it suggests a clear path for determin-

ing whether a population's neutral variation is governed primarily by drift or by draft.

3

SELECTION COEFFICIENTS IN HIV EVOLUTION

3.1 INTRODUCTION

The human immunodeficiency virus (HIV) is a group of lentiviruses descended from several simian immunodeficiency viruses (SIVs). HIV has colonized humans on several occasions. A transmission event from chimpanzees in the early 20th century [39, 125] gave rise to the HIV-1 M clade, which is responsible for the majority HIV infections worldwide.

HIV is of considerable importance due to its role in causing acquired immune deficiency syndrome (AIDS). A typical HIV infection proceeds as follows. Several weeks after exposure, an infected individual experiences an acute flu-like illness as the number of virions swells to around 10^9 individuals [98]. The immune system recovers and hunts down but does not completely eradicate the virus. One of the virus' major hosts during this period is CD4+ cells. CD4+ cells, like most other cells, display fragments of viral protein on their cell membranes. These fragments, known as epitopes, are recognized by cytotoxic T lymphocytes (CTLs), whose major histocompatibility (MHC) molecules interact with the epitope. The CTLs induce apoptosis in the infected cells. As months become years, this process gives way to attack by antibodies. The viral load slowly rises until, a decade or so after the initial infection, the infected individual's immune system is severely compromised.

HIV has also attracted significant attention from population geneticists due to its rapid evolution. In terms of sequence divergence, the equivalent of ten million years of fly evolution can occur within one year of HIV infection. Samples from HIV patients therefore provide a rare combination of high levels of divergence and an observable "fossil record" of sequence data.

During acute infection, CTL attack is the primary driver of natural selection in HIV. Individual virions with mutant epitopes that are more difficult to recognize are more successful: such mutations typically carry an intrinsic fitness cost [42, 47, 77, 115], but this can be compensated for by the ability to avoid CTL killing [5, 47]. This difference between fitness cost and avoided killing is known as the *escape rate* [86]. Escape rates are of interest not only for population geneticists but also for immunologists and virologists. The environment that imposes selective constraints on the virus is, after all, the host's immune system; therefore, the strength and time variation in the virus' evolution inform us about the ferocity and possible time

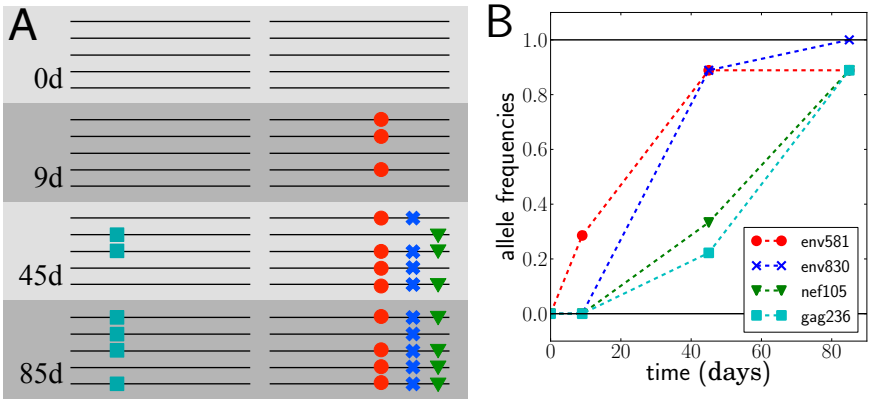

Figure 3.1: A sketch of the behavior of several escape mutants in patient CH58: see also the actual data in Goonetilleke et al. [55].

attenuation of the immune system's response. Moreover, because the escape rate determines the success of individual virions relative to the population average, it is a selection coefficient in the classical sense.

My goal in this project has been to determine, based on the very sparse data available at the time, the escape rates of epitope mutations in acute HIV infection. I begin by summarizing the state of the art and explaining the shortcomings of previous attempts to infer escape rates. I then set forth an alternative model, including a complete but simple computational framework for accurately inferring escape rates even from sparse data.

3.2 MODELS OF IMMUNE "ESCAPE"

Most sexually transmitted HIV infections are effectively founded by a single individual, giving rise to a homogeneous initial population [65, 110]. However, mutations quickly accumulate as the virus reproduces. Mutations that are absent from early samples and either fixed or present at high frequencies later are a hallmark of HIV time series data, as seen in 3.1: see also Goonetilleke et al. [55] and Salazar-Gonzalez et al. [110].

From an evolutionary point of view, there is nowhere near enough time for such mutations to have spread through the population due to drift in the sampling intervals under consideration (weeks or months). The time to fixation for a neutral allele is on the order of N generations. In HIV, N can be as high as 10^9 virions but, for most of acute infection, tends to be closer to 10^7 [15, 25, 99], and a viral generation is on the order of one day [85]. If such mutations are under selection, then their fixation over short time scales is not surprising. The time for a beneficial mutation of selection coefficient s to sweep through the population is on the order of $s^{-1} \log Ns$ generations; refer back to Section 2.1.1.

HIV time series samples such as the ones provided by Goonetilleke et al. [55] and Salazar-Gonzalez et al. [110] are neither particularly frequent nor particularly deep, severely limiting our ability to infer the underlying evolutionary dynamics. The state of the art for inferring the escape rates of sweeping mutations in acute infection has been to fit a deterministic model to the observed mutation frequencies by minimizing some objective function, such as a likelihood or the sum of squared deviations from the model [48, 74, 116]. A good model will respect the underlying population dynamics of HIV. A great model would be robust even in the limit of rare or shallow samples. In the forthcoming pages, I will outline the model I developed for the analysis of HIV time series data from patients CH40, CH58, and CH77 from Goonetilleke et al. [55].

3.2.1 *Independent sites: logistic model*

The simplest possible model considers a beneficial mutation with selection coefficient s growing exponentially in the population. This occurs when, from generation to generation, the expected growth in its frequency ν due to selection is larger than the size of fluctuations due to drift, a condition that obtains roughly when $N\nu > 1/s$, or the mutation drifts to more than $1/s$ individuals. In the infinite sites limit, i.e., assuming that a beneficial mutation arises exactly *once*, this process takes $\sim 1/s$ generations on average.

Suppose one samples a fairly small number of individuals from a large population. Then a beneficial mutation, conditioned on the fact that it is observed in the sample, has certainly passed this drift barrier, and selection dominates its dynamics. The change in its frequency can therefore be written

$$\frac{d\nu}{dt} = s\nu(1-\nu), \tag{3.1}$$

which has the solution

$$\nu(t) = \frac{\nu_0 e^{st}}{1+\nu_0(e^s t - 1)}. \tag{3.2}$$

This model has two parameters, the initial frequency ν_0 and the selection coefficient s. The growth of a fit lineage is exponential at first, then tapers off as the mutation saturates in the population. Close to saturation, this deterministic model is inappropriate: at frequencies close to $1 - 1/s$, drift dominates the allele dynamics once more. Note that the initial frequency ν_0 can be fixed as a parameter by using a suitable population genetic approximation such as the transition from stochastic to deterministic behavior occurring at $\nu_0 \approx (Ns)^{-1}$, but then t in the exponent needs to be $t - \tau$, with τ the transition time.

This model is precisely the one used to infer the selection coefficients of sweeping mutations in acute HIV infection by Ganusov et al. [47]. The initial frequency and selection coefficient most consistent with the observed frequencies of escape mutations are estimated via nonlinear least-squares regression: in the logistic function, the selection coefficient and initial frequency are treated as free parameters. Stochastic effects are ignored, as the allele frequencies observed are quite far from either drift barrier.

Selection coefficients on the order of 0.1 are inferred, with later escapes tending to be much weaker than earlier ones. One possible explanation for this decrease in escape rate is that earlier sweeps appear stronger because the virus is under more vigorous attack, so avoided killing plays a stronger role in determining the escape rate. Potential mechanisms for this attenuation include competition between epitope-specific CTLs for resources such as cytokines or for binding sites, as well as CD8+-mediated control of viral replication. Another possibility is that later escapes incur a higher fitness cost.

The model depends on the assumptions of large population size (so that drift is ignored), no recurrent mutations, and independence between sites. Large population size is a reasonable assumption, as $N \approx 10^7 \gg 1/s$. A more realistic model incorporating recurrent mutations,

$$\frac{dv}{dt} = sv(1-v) + \mu(1-2v), \tag{3.3}$$

is an unnecessary complication, provided $\mu < s$. There is good reason to think this is the case, as often $s \approx 0.1$, whereas $\mu \approx 10^{-5}$ per site per generation; we present the model here for completeness and because it becomes useful later. The solution is

$$v(t) = \frac{1}{2s}\left[s - 2\mu + R\tanh\left(\frac{\alpha + t}{2}R\right)\right], \tag{3.4}$$

where $R = \sqrt{s^2 + 4\mu^2}$ and $\alpha = (4\mu - 2s)/(4\mu^2 + s^2)$. It is not hard to see that the logistic model re-emerges if $s > \mu$, as $R \to s$ and $\alpha \to -2/s$: the alternate form $\tanh(x) = (1 - e^{-2x})/(1 + e^{-2x})$ makes this more obvious.

Independence between sites means that the frequencies of different alleles are uncorrelated. This behavior necessitates a high recombination rate, so that recombination can quickly break up correlations. If there is no epistasis, i.e., if the fitness of a multiple mutant is simply the sum of the individual mutations' fitnesses, the fate of an allele will depend only on itself, not the behavior of other alleles.

3.2.2 *Toward a more realistic model*

In natural populations, independence between sites can fail to obtain for realistic values of ρ, especially when selection causes fit alleles

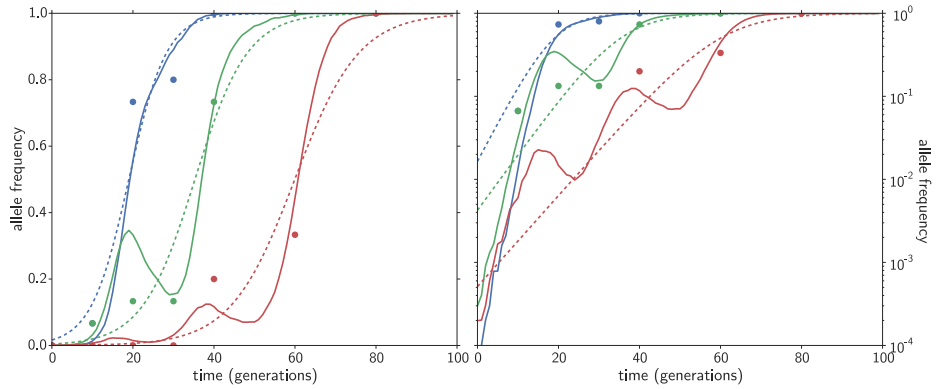

Figure 3.2: Studies such as Ganusov et al. [47] fit observed allele frequencies (dots) against a logistic model (dotted lines). However, if linkage is important, then allele frequencies evolve according to more complicated dynamics (solid lines), and a logistic function will underestimate the escape rate. This is visible in the right plot: when an allele's dynamics transition from drift- to selection-dominance (i.e., when $\nu \sim 1/Ns$), the slope (on a log scale) roughly equals the escape rate. Here the true escape rates are $\{0.3, 0.22, 0.15\}$ but the inferred rates are $\{0.21, 0.15, 0.12\}$.

to expand quickly. The fate of an allele then depends sharply on its genetic background, i.e., on the alleles in its genomic neighborhood. A neutral allele that is linked to an expanding beneficial mutation will ride the wave to high frequency, and an allele present on unfit individuals will be quashed. If selection is strong enough relative to recombination, neutral alleles can remain linked to beneficial genetic backgrounds for the entirety of their lifespan.

A similar pattern characterizes competing beneficial mutations. When beneficial mutations are common and recombination is rare, correlations (negative or positive) between beneficial mutation frequencies are likely to arise. Double beneficial mutants will sweep very quickly through the population, but beneficial mutations that appear in distinct individuals will compete with each other, limiting their ability to sweep. If recombination is too rare to decorrelate them during their solo sweep times, $\sim s^{-1} \log Ns$, they will interfere. In the infinite-sites limit, one variant will go extinct, with the other's sojourn time being extended due to the interference. If we allow for recurrent mutations, then it is possible for both mutations to sweep. But the only way this can happen is for one of the mutations to arise independently on a genetic background containing the other. Thus, if recombination is rare, treating beneficial alleles as independent will cause us to persistently underestimate the selection coefficients of the sweeping mutations, as we will ignore the butting-up of one mutation against another during expansion.

In HIV, recombination does occur but is rare: $\rho \approx 10^{-5}$ per site per generation [8, 90]. While frequent switching of the reverse tran-

scriptase between the two copies of the virus' genome causes recombinants to be produced quite often [75], such recombination occurs *only* between those two copies. The major exception is in cases of coinfection, where two virions infect the same cell. This is a sufficiently rare event that ρ is kept low, so that recombinant beneficial mutants are not likely to arise during the lifetimes of those mutants. As a result, persistent linkage among beneficial mutations cannot be neglected, and a logistic model is inappropriate, as summarized in Figure 3.2.

3.2.3 *Appropriate simplifications for HIV*

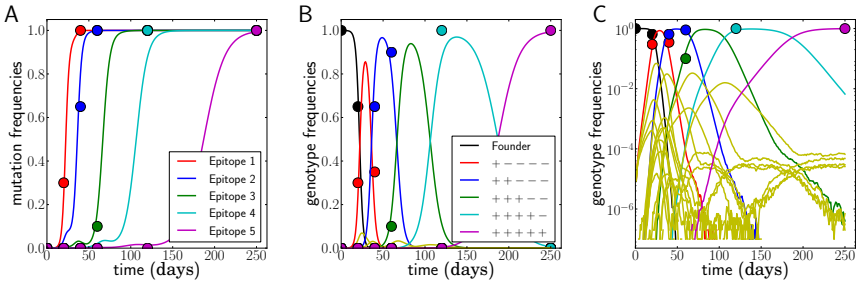

Figure 3.3: Dominance of a handful of genotypes. The dynamics of individual alleles may appear roughly logistic (panel A), but it is more accurate to consider the frequencies of their genotypes (panel B). A model that treats genotypes rather than alleles is rendered tractable by the fact that most genotypes are only ever present at low frequencies and can be ignored (panel C, yellow lines).

A complete model that incorporates the dynamics of L polymorphic loci or 2^L genotypes will not be tractable. However, the proper approximations enable us to write down a model that is still an improvement on the logistic model. In the patient samples we consider, several loci experience beneficial mutations that sweep, but most combinations of these beneficial mutations do not appear in samples at all. Rather, beneficial mutants appear to sweep in a clear order, so that first the wild type is observed, then a genotype containing the first escape mutant, then a genotype containing the first and second, and so on. If L loci experience beneficial mutations during the relevant time interval, then there are $L+1$ such "dominant" genotypes. This situation is outlined in Figure 3.3.

In principle, the genotype dynamics will not be entirely deterministic, meaning they cannot be captured simply by a system of ODEs; mutations and drift are stochastic processes. However, we can use our understanding of population genetics to predict roughly when a mutation should transition from stochastic to deterministic behavior. We refer to these crossover times as *seed times* and treat them as nuisance parameters; we will constrain them by applying population genetic principles.

Furthermore, escape mutants that appear and establish early in infection will affect the dynamics of later mutants, if nothing else by constraining their seed times. But by the time later mutants become significant in the population, earlier ones are already at high frequency. This means that, when the time comes to maximize the likelihood of our model, we will not need to bother with fitting the escape rates and seed times of all L escape mutants simultaneously. Instead, we can fit the escape rate and seed time of one mutant at a time, adding information from later escapes into our model in a stepwise fashion.

We can now begin to outline our model more rigorously. Consider an arbitrary genotype g_j containing escape mutants at locus 1 through j. Suppose the sampled frequency k_{ij} of genotype g_j at time point t_i is drawn from a population in which the true frequency of g_j is $\gamma_j(t_i)$. Then the probability of observing a particular sample is

$$P(\text{sample}) = \frac{n_i!}{\prod_j k_{ij}!} \prod_j \gamma_j(t_i)^{k_{ij}}, \tag{3.5}$$

where n_i is the sample size at time point t_i. The challenge is to select the best values of $\gamma_j(t_i)$ for all the observed genotypes.

In general the frequencies $\gamma_j(t_i)$ will depend on the escape rates at individual loci via their individual growth rates,

$$F(g,t) = F_0(t) + \sum_l \epsilon_l h_l, \tag{3.6}$$

where h_l is 1 for an escape mutant at locus l and 0 otherwise and ϵ_l is the escape rate (selection coefficient, usually written as s) of that mutation. $F_0(t)$ is a time dependent modulation of the total growth rate: we fix it so that the population size remains constant, i.e., the average growth rate is zero. It could in principle be modulated according to external constraints, such as variable numbers of target cells [46, 100].

It may appear that we ignore the effects of epistasis, by simply summing the individual effects of each escape mutant and ignoring possible higher order contributions. However, the ϵ_j we ultimately fit are not completely independent of each other: rather, they are the ϵ_j conditioned on a particular genetic background containing mutations $1 \ldots j-1$, which are the only escape rates that can really be estimated given the available data. Our method is agnostic as to how important epistasis actually is.

Provided that deterministic factors contribute much more to $\gamma_j(t)$ than stochastic ones do,

$$\frac{d\gamma_j(t)}{dt} = F(g_j,t)\gamma_j(t) + \mu\left[\gamma_{j-1}(t) - \gamma_j(t)\right] \tag{3.7}$$

is a good summary of $\gamma_j(t)$'s dynamics. (For γ_0, i.e., the wild type, the first term in the mutation contribution does not apply.) This obtains

when the change in $\gamma_j(t)$ due to selection is more important than that due to drift and the change in $\gamma_j(t)$ due to mutations $\mu\left[\gamma_{j-1}(t) - \gamma_j(t)\right]$ is large enough to be essentially deterministic. These amount to a constraint on the size of $\gamma_j(t)$ and $\gamma_{j-1}(t)$.

This constraint can be realized by selecting an appropriate seed time. Genotype g_j arises at a rate $\mu N \gamma_{j-1}$ from the g_{j-1} genotype. This is unlikely to happen while γ_{j-1} is small, but once it is appreciably large, g_j is produced frequently. Note that we ignore the establishment probability $\sim 1/\epsilon_j$ in our analysis, as we assume that successive escape mutants generally arise many times and that ϵ_j is rather high: the establishment time is therefore of order 1. (We also implicitly set the generation time to 1 day: if it happens to be two days, this has effects similar to cutting the population size in half.) The distribution of the seed time τ_j is determined by the rate at which g_j is produced and the probability that it has not been produced up to this point: the latter is given by $\exp(-\mu \int_{\tau_{j-1}}^{\tau_j} N\gamma_{j-1}(t)dt)$. Accordingly, the distribution can be approximated by

$$Q(\tau_j|\gamma_{j-1}(t)) \approx \mu N \gamma_{j-1}(\tau_j) e^{-\mu \int_{\tau_{j-1}}^{\tau_j} N\gamma_{j-1}(t)dt}. \tag{3.8}$$

This limits the plausible values for τ_j based on γ_{j-1}. In conjunction with the observed frequency counts, we should be able to reasonably estimate the ϵ_j for each escape mutation. Since the trajectories γ_j are uniquely specified by the ϵ_j and τ_j via Equation 3.7, we can write down the full likelihood function:

$$\mathcal{L}(\{\epsilon_j, \tau_j\}|\Theta) = P(\Theta|\{\epsilon_j, \tau_j\}) \propto \prod_i P(\text{sample}_i|\Theta) \prod_j Q(\tau_j|\Theta) U(\epsilon_j). \tag{3.9}$$

Here, $U(\epsilon_j)$ is a prior that biases our escape rate estimates toward lower values. We select a Laplace prior $U(\epsilon) = \exp(-\Phi\epsilon)$, where higher values of Φ favor lower escape rates. In general the prior makes little difference but slightly slopes our likelihood surface, yielding conservative estimates. For most of our analysis, we set $\Phi = 10$.

3.3 INFERRING ESCAPE RATES

We implement the previously described ODE model and likelihood function in a simple fitting scheme. A brief outline follows.

1. Manually count the number of occurrences of each escape mutant genotype, omitting any that do not reach high frequency.
2. Generate initial estimates of the escape rates and seed times for each of the escape mutations using simple one-dimensional models in a maximum likelihood framework.

3. For each mutant, jointly estimate the escape rate and seed time conditioned solely on the behavior of previous mutants, in a maximum likelihood framework using our ODE model.

4. For each mutant, adjust escape rate and seed time estimates while keeping the trajectories for all other mutants fixed. Repeat until convergence.

5. Generate posterior distributions by Markov chain Monte Carlo.

For our initial estimates, we note that the first mutant genotype is seeded from a wild type population that quickly reaches around 10^9 individuals [25, 99]. Accordingly, we rely on Equation 3.4 to estimate ϵ_1, as the injection of mutant individuals even early in the growth of the mutant lineage is significant: see Ganusov et al. [48]. We assume that $\tau_1 \approx 20$ days prior to the date of first sampling, as each of the patients was identified in Fiebig stage II [1], corresponding roughly to the onset of symptoms [43]. Later escape rates are initially estimated by using Equation 3.2, with the seed time τ_j and escape rate ϵ_j jointly fitted.

The best fit is determined by maximizing the likelihood function Equation 3.9 using the `scipy` function `fmin` [96]. For the initial estimate phase of our analysis, the data-dependent portion of the likelihood is given by

$$P(\text{sample}_i|\Theta) = \sum_{j=1}^{L} \binom{n_i}{k_{ij}} \nu_j(t_i)^{k_{ij}} (1-\nu_j(t_i))^{n-k_{ij}} : \tag{3.10}$$

that is to say, the observed allele counts ν_j are assumed to be binomially sampled from an underlying frequency distribution described by our logistic(-like) functions. Here, k_{ij} is the observed number of counts of allele j at time point i. Note that the sum begins at 1: there is no sense in fitting the "wild type" via a logistic.

Next, we iterate the procedure by adjusting the seed time and escape rate of the first locus only. The data-dependent portion of the likelihood becomes

$$P(\text{sample}_i|\Theta) = \sum_{j=0}^{1} \binom{n_i}{k_{ij}} \gamma_j(t_i)^{k_{ij}} (1-\gamma_j(t_i))^{n-k_{ij}}, \tag{3.11}$$

where the γ_j obey equation Equation 3.7 and the k_{ij} are now the counts of genotype g_j. The computationally expensive numerical solution of these ODEs is implemented in `C++` and exposed in Python using `SWIG` [9]: the sum ranges from 0 (wild type) to 1 (mutant genotype with an escape at the first locus only). We reiterate this procedure for subsequent escapes up to L, so that the data-dependent portion of the likelihood function becomes

$$P(\text{sample}_i|\Theta) = \sum_{j=0}^{L} \binom{n_i}{k_{ij}} \gamma_j(t_i)^{k_{ij}} (1-\gamma_j(t_i))^{n-k_{ij}}. \tag{3.12}$$

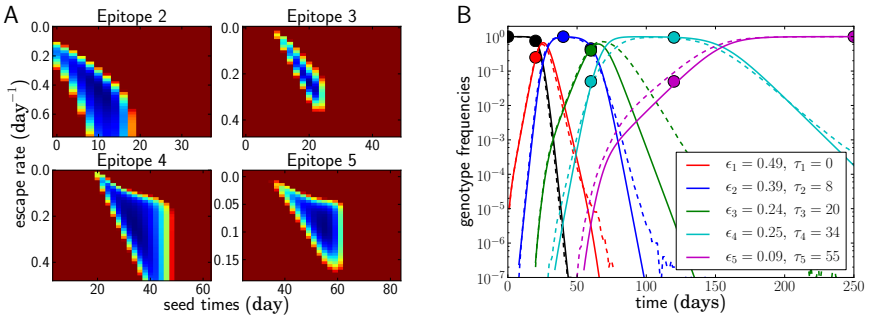

Figure 3.4: Fitting of epitope frequencies from toy data. We fit the first epitope's escape rate with τ fixed, then successively add later escapes by minimizing the likelihood across their two-dimensional likelihood surfaces (panel A). Subsequent rounds of fitting are performed by adjusting ϵ and τ for each epitope, keeping later escapes fixed In this example, $N = 10^7$, $\mu = 10^{-5}$, and $\epsilon \in [0.5, 0.4, 0.25, 0.18, 0.08]$ per day. The estimated ϵ values (panel B) are close.

(The only change from Equation 3.11 is the upper summation bound.)

Selection of the best fit of ϵ_j and τ_j is accomplished thusly: We construct two dimensional likelihood surfaces, with ϵ_j on one axis and τ_j on the other, and minimize the likelihood along this surface. All previous ϵ_k and τ_k ($k < j$) are held fixed. Minimization is somewhat confounded by the fact that τ_j is discrete by construction: there are no "fractional" generations in our approach. Therefore, minimization algorithms such as gradient descent, which rely on a continuous likelihood surface, are not appropriate, as likelihood surfaces appear locally flat in the τ_j direction.

The minimization algorithm we employed is a custom greedy algorithm: we permit vertical, horizontal, or diagonal steps on the surface with size $\delta\tau = \pm 1$ day and $\delta\epsilon = \pm 0.02$ per day. This is repeated until no favorable move is found, at which point $\delta\epsilon$ is adjusted to ± 0.01 and ± 0.001 per day. On the rare occasions where two or more mutations were present at exactly the same frequencies in the population (see Section 3.3.2), we assumed that their selection coefficients were equal and permitted steps that changed the selection coefficients, individual seed times, all seed times, or random combinations of seed times and selection coefficients.

The use of a greedy algorithm is justified by the observation that two-dimensional slices of the likelihood space overwhelmingly tend to be smooth, with a single, wide valley. This is intuitive. As ϵ_j decreases, the data-dependent portion of the likelihood function favors lower values of τ_j, because establishment needs to occur earlier in order for sweep times to agree with the data, but τ_j is also dependent on the fitted dynamics of the $j-1$ genotype, as g_j cannot arise while

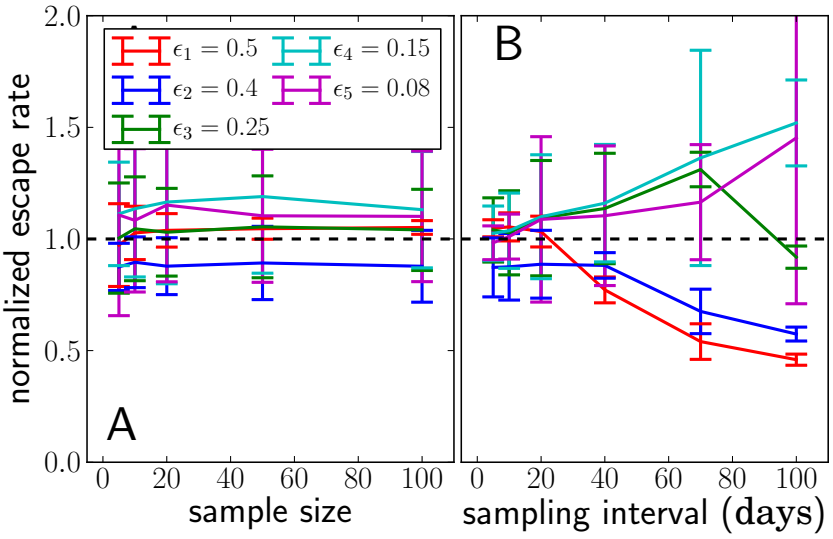

Figure 3.5: The effects of sampling uncertainty due to shallow (panel A) or rare (panel B) sampling. Imputed escape rate estimates are very noisy when sampling is infrequent, as mutants are sampled essentially only at frequencies 0 and 1. In contrast, even shallow sampling is not a problem for our maximum likelihood model. The dotted line indicates an unbiased estimate, i.e., estimates are normalized by dividing by the true escape rate.

g_{j-1} is too rare. These competing constraints suggest that it is not reasonable to expect a complex likelihood surface with many valleys.

Once each of the ϵ_j and τ_j has been estimated using this approach, we reiterate the fitting process, this time incorporating data from later escapes and keeping all other escape rates and seed times fixed. This procedure is repeated until all escape rates and seed times converge, which takes about one minute on one 2011 desktop machine (Apple iMac i7 2.93GHz).

To map the likelihood surfaces and thereby assign confidence intervals to our estimates, we attempt to change all seed times and escape rates by $\delta\tau = \pm 1$ day and $\delta\epsilon = \pm 0.01$ per day with random sign. Steps are accepted with probability $\min(1, \exp(\Delta))$, with Δ the difference in log-likelihood before and after the change (so that steps improving the likelihood were always accepted). The resulting Markov chain is sampled every 1000 moves. This step takes about 20 minutes on the same desktop computer.

3.3.1 *Testing our inference method*

To demonstrate the accuracy of our method, we relied on *in silico* data where it was possible to know the escape rates *a priori*, as well as control for parameters such as N and μ, which regrettably it is hard to estimate with a high level of precision in natural populations. The

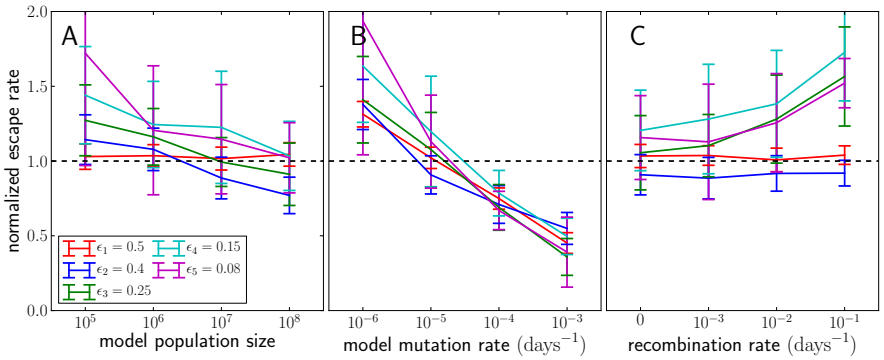

Figure 3.6: The effects of faulty model parameter estimates. We simulated escape dynamics with fixed, known values of $N = 10^7$, $\mu = 10^{-5}$. In A and B, model estimates of N and μ are as shown. In the simulations considered in panels A and B, $\rho = 0$; in panel C, it varies as shown. If estimates of either N or especially μ are too high, successive escape mutants are impute to arise too early, deflating the escape rate estimates. Underestimates of ρ primarily bias the escape rate estimates for later escapes by allowing them to be seeded earlier than we permit. When N, μ, and ρ are consistent between model and simulation, the estimates are unbiased.

data were generated with the forward population simulation package `FFPopSim` [128]. We initialize a population at fixed size N and mutation rate μ, setting the selection coefficients of $L = 5$ mutations to $\{0.5, 0.4, 0.25, 0.15, 0.08\}$. We record the genotype frequencies at variable sampling intervals; we then sample binomially from the genotype frequencies. The escape rate values were chosen because they are about what we expect to observe in HIV, and the selected sampling intervals mimic the sparseness of data from HIV time series samples.

It is important to recognize two possible sources of error in our inference. The first is uncertainty due to limited sample size or frequency. The second is systematic error due to an inappropriate choice of model parameters, such as N and μ, or even an inappropriate model altogether. We use our toy data to probe the effects of both.

More frequent or deeper sampling should generally reduce the uncertainty in our parameter estimates. These conditions would make it possible to determine the mutation's escape rate simply from the rate at which it increases, which is how studies such as Ganusov et al. [47] attempt to proceed. However, in the sparse data typical of longitudinal HIV studies, it often happens that a mutation is observed at zero frequency and at fixation, with no intermediate frequencies observed. We find that even when sample sizes are kept quite small, the constraint we impose on escape rates via constrained seed times and our Laplace prior forces escape rate estimates into an accurate range, even with less data. On the other hand, infrequent sampling can bias estimates significantly, as shown in Figure 3.5.

Systematic error due to a poor choice of model or parameters is a more pressing concern. We first investigated the effects of faulty estimates of N or μ. We forward simulated populations with the aforementioned escape rates, taking samples of size 20 at time points $t = \{0, 20, 40, 60, 120, 250\}$ generations, and let the simulated $N = 10^7$ and $\mu = 10^{-4}$ or 10^{-5}. We then attempted to infer escape rates given values of N ranging from 10^5 to 10^8 and values of μ ranging from 10^{-6} to 10^{-3}.

Results are shown in Figure 3.6. As it happens, our estimates depend only weakly on the precise values of N and μ: they are robust against errors in these parameters. One reason is that the seed time prior has a relatively sharp peak: the log likelihood depends on the log of $N\mu Q(\tau_j|\gamma_{j-1})$, which peaks when $N\mu\gamma_{j-1} \approx 1$. In the early stages of growth, γ_{j-1} is expanding exponentially, so it dominates the seed time prior: the $N\mu$ prefactor influences the seed time only logarithmically. A notable exception occurs when $N\mu$ is small. In this case, the seed time prior is no longer sharply peaked: the dynamics are highly stochastic, so seed times can take on a wide range of plausible values.

In fact, even though we have explicitly assumed no recombination in our model, our estimates are nonetheless robust even when recombination is introduced in simulations. If the rate of recombination is low, then its major effect is to cause mutations to "seed" slightly earlier. Therefore, by wrongly assuming no recombination, we infer that mutations seed later than in reality, slightly skewing their escape rates toward higher values. In general the flexibility afforded by probabilistically constraining the seed times allows our estimates to remain rather accurate even if recombination is introduced. Recombination does not induce significant errors until ρ is of order μ or higher, which implies a very high rate of coinfection: otherwise, mutations are much more important in determining the seed time than recombination is, and we are safe in ignoring it.

3.3.2 *Unobserved intermediates*

In time series samples, the dominant genotype at two subsequent time points sometimes differs by more than one escape mutation. There are, broadly speaking, two possible explanations for this. One is that several beneficial mutations arose in rapid succession on the same genetic background. The other is that the mutations interact via sign epistasis: the first mutation carries a significant fitness cost, which is undone by a compensatory mutation. Both possibilities can be accounted for in our scheme.

In the case of individually beneficial mutations that are always observed together, intermediate genotypes can be incorporated into our model by assuming, conservatively, that the intermediates have the

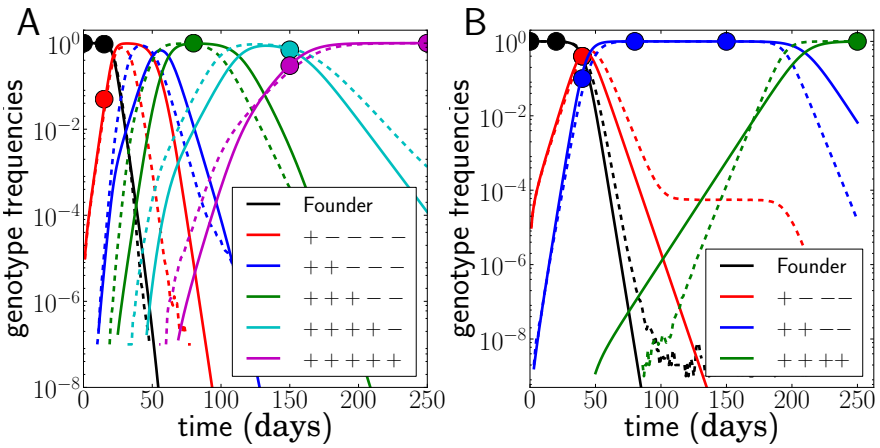

Figure 3.7: The effects of unobserved intermediates or compensatory mutations (valley crossing). In A, the blue genotype is not sampled even though it was transiently dominant. We assume the blue mutant has the same escape rate as the green one for parsimony's sake. In B, the third and fourth escape mutations are deleterious unless they appear together: accordingly, the green genotype must appear as a double mutant from the blue one, and seed time estimates are broadly distributed.

same escape rate. This keeps the number of parameters down, though the resulting escape rate estimates should be taken with a grain of salt: there simply is not enough information available to estimate more than an average escape rate for such intermediates. Somewhat surprisingly, escape rate estimates tend to *increase* as more intermediates are assumed. This is intuitive if one considers the seed time priors: each intermediate must be at a high enough frequency to make the seeding of subsequent mutants likely.

Deleterious intermediates are somewhat more difficult to incorporate. Broadly speaking, there are two ways to generate a double mutant when intermediates are deleterious [124]. First, deleterious intermediates can segregate at low frequencies due to drift, with double mutants occasionally arising on a deleterious lineage, a process known as tunneling. Second, multiple mutations can arise on one individual.

We consider only the latter. The disadvantage s of an intermediate can be quite significant, meaning that it is unlikely to drift to high frequencies. In general, a single deleterious intermediate can drift to frequency $\nu = 1/Ns$, or about $1/s$ individuals, before selection begins to sharply limit its growth. These $1/s$ individuals will birth double mutants at a rate μ/s, which is quite small. Moreover, the product $N\mu$ is quite large, and in fact $N\mu^2$ is often not much less than one. This means that the deleterious intermediate gap can plausibly be crossed within one step, i.e., a double mutant can appear all at once on a wild type background within a realistic length of time.

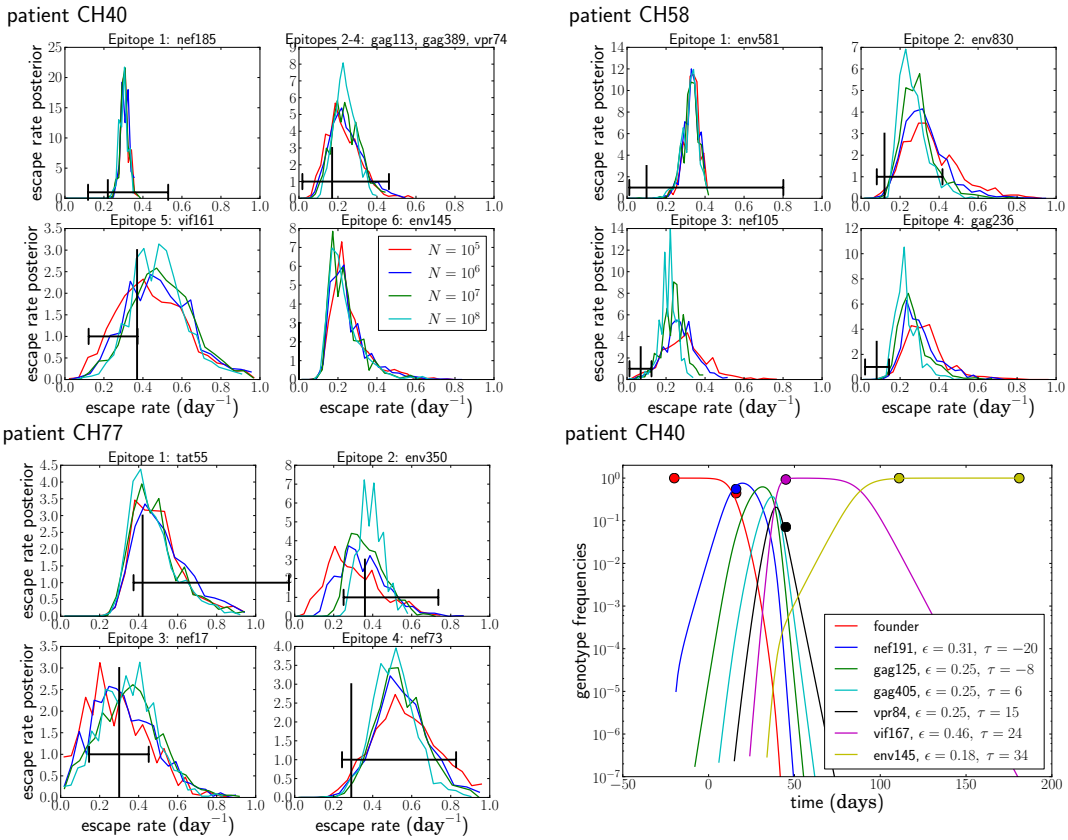

Figure 3.8: Escape rate estimates for patients CH40, CH58, and CH77, assuming no recombination and $\mu = 10^{-5}$. The curves show escape rate posterior estimates generated by MCMCMC: black lines indicate the estimates and confidence intervals given by Ganusov et al. [47]. The lower right shows the most likely trajectory for patient CH40, with $N = 10^7$.

Thus, in the simplest case where intermediates are all deleterious, the only change that really needs to be made is to replace μ with μ^{k+1} in the seed time prior, where k is the number of deleterious intermediates, in the seed time prior. In more complex escape scenarios, the rate at which double mutants appear can be calculated using a branching process approximation. Unfortunately, in the simple case, the seed time prior turns out to not to be very well constrained. If $N = 10^8$ and $\mu = 10^{-5}$, the width of the seed time prior (dependent on $N\mu^2$) is more than 100 days. More data are required in such a scenario in order to accurately infer the escape rate.

3.3.3 *Escape rates in patient data*

We proceed to stimate the escape rates of beneficial mutants in patients CH40, CH58, and CH77. CTL escape in these patients was previously characterized by Goonetilleke et al. [55] and Salazar-Gonzalez et al. [110] and analyzed mathematically by Ganusov et al. [47]. These

time [days]	founder	env581	env830	nef105	gag236
9	5	2	0	0	0
45	0	0	5	3	0
85	0	0	0	0	8

Table 3.1: Format of input data for patient CH58. Escape mutations are ordered first by the time at which they first appear in samples and then by their adundance. The value in each cell is the number of times a sequence is observed at that time point containing that column's escape mutation and all previous ones.

patients were untreated, so their viral populations evolved in a fairly pristine manner. The sequences were obtained via single genome amplification followed by traditional sequencing. Unfortunately, the samples are shallower ($5-10$ samples per time point) and less frequent than the toy data we used to test our model, so our escape rate estimates are somewhat imprecise. Furthermore, we do not know exactly when each patient was infected or when CTL selection began: the time stamps given in the data are relative to the first sample drawn. We estimate that both of these began roughly $\tau = 20$ days before the first sample date, consistent with a rough timeline of HIV symptom progression [43], as each patient presented in Fiebig stage II [1]. Changing the sampling time to $\tau = 0$ days affects estimates of the escape rate at the first locus but not subsequent loci.

Fixed parameter estimates for our model proceeded as follows. In chronically infected patients, there are around 4×10^7 infected cells [57], but during acute infection, this can vary substantially, as the viral load can be much higher during peak viremia, as well as due to any number of immune-related factors. Consistent with our testing procedure for the toy data, we determined posterior distributions for N ranging from 10^5 to 10^8. The per-site mutation rate in HIV is around 10^{-5} [84], but epitopes may permit escape at more than one site: in this case, a mutation rate closer to 10^{-4} is more appropriate. We consider both possibilites. We ignore recombination for reasons previously discussed: stepwise mutations are anyway likely to be more important than recombination in generating multiple mutants.

For each of the three patients, we considered all nonsynonymous mutations that are eventually sampled at high frequency as possible escape mutants. We treated nearby mutations in the same epitope as part of the same escape: a more complete treatment might regard each of these as a distinct competing escape. We further refined this by considering only time points early in infection, with more than 5 samples per time point, and only the first handful of strong escapes: later in infection, nonsynonymous diversity is very high, and our approach is ill suited to fitting a time course for many of the mutations observed. Mutations whose frequencies do not increase monotonically, such as

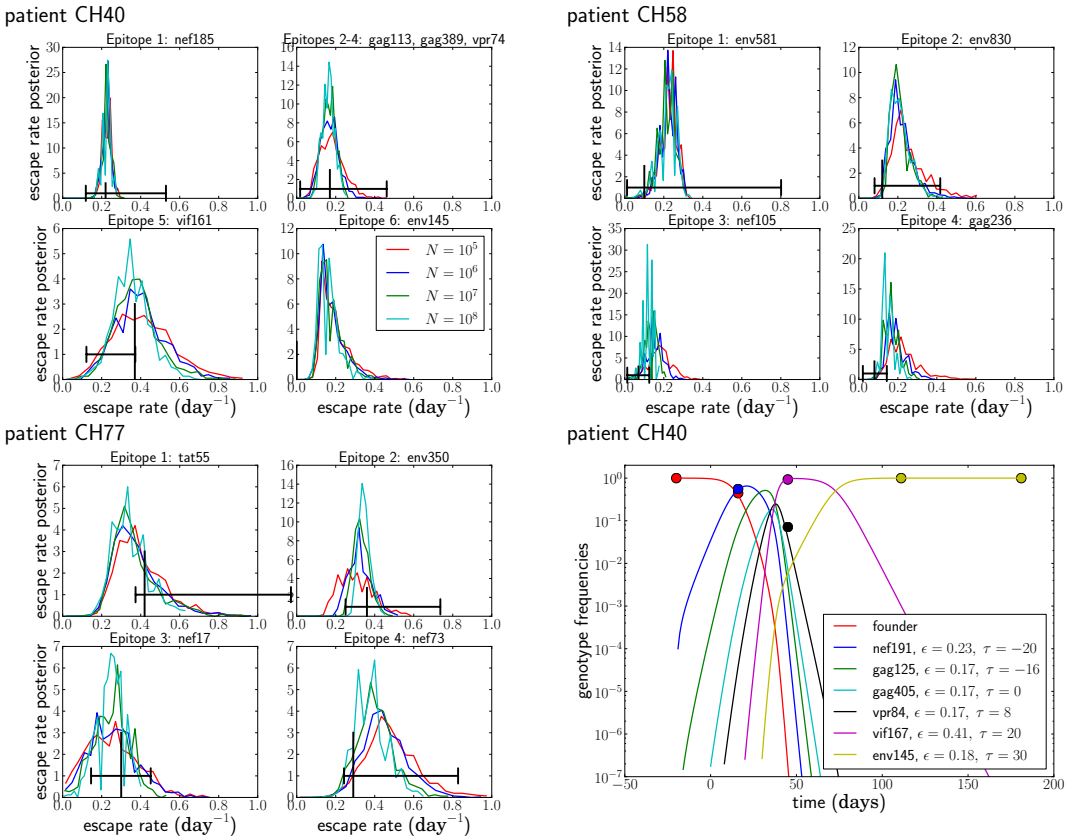

Figure 3.9: Escape rate estimates for patients CH40, CH58, and CH77, assuming $\mu = 10^{-4}$ instead of 10^{-5}. Simulation parameters are otherwise as in Figure 3.8.

pol80 in subject CH40, were ignored. Mutation frequencies and their linkage into multi-locus escape genotypes were imputed based on the alignment of Salazar-Gonzalez et al. [110]. This regrettably omits linkage between the 5′ and 3′ halves of the genome, but that can be imputed by assuming sequential escapes, as we do. The number of mutant genotypes was counted and input into a tab-delimited table, as represented in Table 3.1.

In CH40, the first five time points were $t = 0, 16, 45, 111, 181$ days. We identified six epitopes as escape targets: nef185, three indistinguishable escapes in gag113, gag389, and vpr74, and two escapes in vif161 and env145. The number denotes the beginning of the 18-mer peptide encompassing the epitope. The env145 mutation is one that Ganusov et al. [47] did not analyze: the 145 refers to the mutated amino acid in gp120. We treated gag113, gag389, and vpr74 as sequential, individually beneficial escapes, as discussed previously (they are assumed to have the same escape rate).

The dynamics were somewhat simpler in the other two patients. In CH58 we considered samples drawn at the first four time points ($t = 0, 9, 45, 85$ days) and identified four escapes: env581, env830, nef105, and gag236. In CH77 we considered only the first three time points

($t = 0, 16, 32$ days) and identified four escapes: tat55, env350, nef17, and nef73.

We proceeded to estimate escape rates as previously outlined. For each escape mutant, we used a simple one-dimensional logistic (or nearly logistic) model to obtain an initial estimate for the escape rates ϵ_j. We refined this by iterating over our multi-epitope fitting process a maximum of five times (after which convergence obtained). We then performed an MCMC simulation using our likelihood function, letting the Markov chain run for 10^6 steps and sampling every thousand steps.

Our estimates are displayed in Figure 3.8 and Figure 3.9 for $\mu = 10^{-5}$ and 10^{-4}, respectively, with the best fit model for patient CH40 and the estimates of Ganusov et al. [47] in black. Although we obtain fairly broad posterior distributions, we nonetheless suggest that escape rates are higher than previously estimated. In contrast to previous work, we do not infer a strong correlation between the time at which an escape mutant emerges and its escape rate. We suggest that this is an artifact of the underlying population genetics: escapes that happen to be strong establish more easily. It is worth noting that, for the multiple mutants, our estimates are especially noisy: still, strong selection is clearly required for multiple mutants to fix in just a few weeks.

3.4 CONCLUSION

We have developed a method for inferring viral escape rates from serially sampled sequences, based on a realistic understanding of the behavior of rarely sexual populations such as HIV. Our method capitalizes on the sequential nature in which new mutations arise, which allows us to respect linkage between beneficial mutations without excess nuisance parameters or complication. Typical methods for inferring escape rates have proceeded by fitting the observed data to a logistic curve. This has been used to analyze the dynamics of recombinant HIV [81], drug resistance [16, 97], and CTL escape [4, 5, 46, 47, 101]. These methods can be useful if data are deeply and frequently sampled enough, but when data are sparse, they provide only lower bounds. The implicit assumption in such models is that the recombination rate is high, but recombination in HIV is somewhat rare [8, 90].

Because our method respects interference between competing escape mutants, it tends to estimate somewhat higher escape rate values than prior methods. Even with a large prior against high escape rates ($\Phi = 10$), we estimate that ϵ for the first few escapes can be as high as $0.3 - 0.4$ per day. In fact, provided the population size is large, these values are essentially insensitive to the choice of prior. Early in infection, $N = 10^7$ is a reasonable estimate [15, 25, 99]. But if the

population size is even smaller, then relaxing the prior against high escape rates results in even higher estimates. Clearly it is important to consider competition between escape mutants in order to avoid underestimating the strength of selection acting on them.

Like any method for inferring escape rates, ours is somewhat limited by the sparse data available. For example, our escape rate estimates for the three simultaneous escapes in patient CH40 have broad posterior distributions: with identical dynamics and essentially no time resolution, our estimates are accurate but not particularly precise. More precise results, especially correlations with factors such as immunodominance [80], will necessitate more data.

As previously mentioned, we do not observe a strong correlation between escape rates and the time at which a mutation appears. Rather, our model suggests an explanation for the phenomenon of slower sweeps following later escapes other than possible time attenuation or higher fitness costs later in infection. The population size and mutation rate of HIV are large enough ($N\mu \gg 1$) that every single nucleotide variant appears somewhere in the population every generation. Therefore, the escape mutants that sweep early are naturally the ones with the strongest escape rates, since they have the lowest drift barrier to overcome and outcompete other mutations upon crossing the barrier. Selection does not significantly affect subsequent mutants until they appear on a genetic background containing the earlier, stronger escapes.

This method for analyzing time series data involving many sweeping loci could be extended to other situations where a population can be observed and sampled over relevant time scales, such as cancer evolution or microbial evolution experiments. Many rapidly adapting populations in which multiple beneficial mutations segregate simultaneously are of great clinical significance. A realistic attempt to understand their behavior must respect the role of interference and linkage in determining their evolutionary trajectories.

4

GENETIC DRAFT AND THE SHAPE OF GENEALOGIES

4.1 INTRODUCTION

The genealogy of a population is a summary of its evolutionary history, which ultimately is what determines the patterns of genetic diversity we observe. In fact, if the population is asexual, a genealogy can be a very accurate model of its history: each individual has exactly one ancestor in the previous generation. In Chapter 2, I argued for a close link between the shape of a genealogy and the processes that have given rise to it. But what, exactly, can genealogies tell us about those processes? Very rapid adaptation can yield genealogies more consistent with the BSC than with the Kingman coalescent [89], but processes such as demographic expansion or bottlenecks can also distort genealogies [129]. Can we disentangle these effects and zero in on rapid adaptation?

That question is what gave rise to the following project. I begin by outlining standard site frequency spectrum based methods for determining whether a population (or a suitably defined region in a population's genome) is experiencing positive selection. As previously suggested in Section 2.4, there is a tight relationship between a genealogy's shape and the population's site frequency spectrum. Skewed branchings can correspond to a non-monotonic SFS, which in turn suggests that neutral variation is shaped by hitchhiking rather than draft. I elaborate further on this relationship here.

The outcome is a procedure for differentiating between drift and dratft that, unlike previous methods, is unaffected by the confounding effects of demography. We consider only the topology and order of branching events on a tree, not the branch lengths thereof, which allows us to focus solely on the exponential amplification of fit lineages consistent with genetic draft. Similar methods have recently been developed by Li [78] and Li and Wiehe [79], who proposed statistical tests for positive selection based on genealogical topologies. We go a step further by explicitly comparing our statistic, the *partition entropy*, to the value of $N\sigma$ (which characterizes a population's adaptation) and showing how it can be used to infer $N\sigma$ when it is not known.

We test our method by applying it to the HA segment of influenza A subtype H3N2. We find that its genealogies are consistent with high values of $N\sigma$ and a rejection of genetic drift. This further strengthens the call for novel null models of sequence variation in rapidly adapting organisms. Most of the work in this chapter refers only to asexual

populations for which a genealogy is well defined, but in Chapter 5, I will outline how it can be adapted to sexual populations.

4.1.1 *Classical tests of selection*

A variety of methods exist for inferring the strength of selection from sequence data. Typically they are used in sliding window analyses or applied to some well defined linkage block or haplotype. One example is the ratio

$$\omega = \frac{K_a}{K_s} = \frac{dN}{dS} \tag{4.1}$$

of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site. In general, $\omega = 1$ indicates neutrality, $\omega < 1$ suggests purifying selection (as nonsynonymous substitutions are disfavored), and $\omega > 1$ (as nonsynonymous mutations, when present, fix more quickly than nonsynonymous ones, leading to a higher substitution rate). Estimating ω from sequence data is often not easy: complicating factors such as codon bias, a skewed transition-transversion ratio, or multiple substitutions at the same site must be taken into account. Moreover, selection frequently occurs in regulatory regions, where the genetic code is not important.

There are many other methods for detecting selection in sequence data. Moving forward, I will consider only tests for selection that depend on the SFS in a population or sample thereof. These can be given straightforward genealogical interpretations, so they can help us to determine exactly how genealogies can inform the evolutionary process. It is worth noting that compound approaches, which combine SFS- and genetic code-based methods, have been somewhat successful in inferring selection [58].

One well known statistic that depends only on the SFS is Tajima's D [119],

$$D = \frac{\theta_\pi - \theta_s}{\sqrt{\text{Var}\theta_\pi - \theta_s}}, \tag{4.2}$$

where θ_π is the average pairwise difference between individuals and $\theta_s = S/a_{n-1}$ is the number of segregating sites divided by the $n-1$th harmonic number $\sum_{i=1}^{n-1} i^{-1}$. Note that that θ_π can be expressed in terms of the per-site "heterozygosity",

$$\theta_\pi = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i(n-i)s_i, \tag{4.3}$$

with s_i the number of mutations that appear in i individuals.

The denominator $\sqrt{\text{Var}(\theta_\pi - \theta_s)}$ is the standard deviation of $\theta_\pi - \theta_s$ under the Kingman coalescent; it is typically computed [119] as

$$\text{Var}(\theta_\pi - \theta_s) = e_1 S + e_2 S(S-1), \tag{4.4}$$

with

$$e_1 = \frac{1}{a_n}\left(\frac{n+1}{3(n-1)} - \frac{1}{a_n}\right) \tag{4.5}$$

and

$$e_2 = \frac{1}{a_n^2 + b_n}\left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{na_n} + \frac{b_n}{a_n^2}\right), \tag{4.6}$$

in which $b_n = \sum_{i=1}^{k-1} i^{-2}$.

Tajima's D has a ready genealogical interpretation. In the infinite sites limit, S scales as the total branch length of the tree $a_n \sim \log n$, as this (multiplied by the mutation rate) is the expected number of segregating sites. The average pairwise difference θ_π scales as twice the average pair coalescence time, as mutations along both lines of descent from a common ancestor contribute to the pairwise difference.

Typically, a value of D less than zero is taken to imply recent positive selection: long terminal branches, which contribute disproportionately to θ_s, suggest a sharp reduction in diversity in the past, consistent with a recent selective sweep. A value greater than zero implies balancing selection: a long pair coalescence time implies that diversity persists for longer than expected under a neutral model, suggesting that haplotypes are prevented from fixing or going extinct, which occurs in the case of heterozygote advantage or frequency-dependent selection. Recall that, under the Kingman coalescent, the SFS scales as $f(\nu) \sim \nu^{-1}$. Essentially, D picks up on deviations from this neutral spectrum: the heterozygosity corresponds to the intermediate portion of the spectrum, and the total number of segregating sites reflects the total area underneath it.

Tajima's D benefits from being an unbiased estimator (i.e., under the Kingman coalescent its expected value is zero). It has no known analytical distribution under the Kingman coalescent, but it can be approximated with a Beta distribution [119]. Today, the most common approach is to simulate the Kingman coalescent or some similar population model many times, generate an empirical null distribution, and thereby impute a p value. As an extremely rough rule of thumb, a value of D greater than 2 or less than -2 indicates significant deviations from neutrality. Unfortunately Tajima's D is not particularly useful for imputing the *strength* of selection: for example, it cannot be used to estimate selection coefficients. However, in sexual populations, a sliding window analysis can be appropriate for finding regions of the genome that may be under strong selection.

Some other common statistics that summarize the shape of different parts of the SFS include Fu and Li's D and F statistics [45],

$$D = \frac{\eta - a_n\eta_e}{\sqrt{u_D\eta + v_D\eta^2}} \tag{4.7}$$

and

$$F = \frac{\theta_\pi - a_n\eta_e}{\sqrt{u_F\eta + v_F\eta^2}}, \tag{4.8}$$

with η and η_e the total number of segregating sites and the number of singletons, respectively. The u and v terms in the denominator are derived from the variances of the numerator terms (under the Kingman coalescent). Fu and Li's D measures the proportion of non-terminal branch length on the tree, and F compares the intermediate-frequency and singleton portions of the SFS, i.e., the pair coalescence time and total terminal branch length.

The last statistic worth mentioning is Fay and Wu's H [40],

$$H = \frac{\theta_\pi - \theta_H}{\sqrt{\mathrm{Var}(\theta_\pi - \theta_H)}}. \tag{4.9}$$

Here,

$$\theta_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 s_i, \tag{4.10}$$

with s_i defined as above. The i^2 weighting term means that the sum heavily favors mutations at the common end of the SFS. The variance is difficult to compute, but note that $\theta_\pi - \theta_H = 2(\theta_\pi - \theta_L)$, with

$$\theta_L = \frac{1}{n-1} \sum_{i=1}^{n-1} i s_i : \tag{4.11}$$

as a result, H can be rewritten as

$$H = \frac{\theta_\pi - \theta_L}{\sqrt{\mathrm{Var}(\theta_\pi - \theta_L)}}, \tag{4.12}$$

in which the variance can be computed exactly [130]. H purportedly compares the high frequency and intermediate portions of the SFS.

4.1.2 *The effects of demography*

A major shortcoming of the use of branch length dependent statistics is that they are all subject to the confounding effects of demography, which cannot be overstated. In genealogical terms, a rapid population expansion, perhaps due to recovering from a disaster or colonizing a new enviroment, implies a lower population size and hence higher merger rate in the past. As a result, terminal branches are long relative to internal branches, and the rare end of the SFS tends to be steeper, similar to the case of positive selection: recall Figure 2.4 and Figure 2.5.

One can approximate Tajima's D's dependence on growth rate by considering the shape of Kingman coalescent genealogies. When population size is fixed, the average pairwise distance scales roughly as $2N\mu$ and the number of segregating sites as $2N\mu \log n$ (for large n, i.e., ignoring the distinction between a_n and $\log n$). The latter is seen by considering the total branch length on the tree $\sim 2N \log n$. However, if a population expands rapidly, almost all the branch length on

the tree will be terminal. This means that the average pairwise distance will scale as $4N_e\mu$ (introducing N_e solely for convenience), or twice the total coalescence time for the entire population (ignoring the $1 - 1/n$ term). The number of segregating sites, on the contrary, will scale as $N_e\mu n$, effectively increasing by a factor of $n/2\log n$: each of the n terminal branches will persist for N_e generations, accumulating $N_e\mu$ mutations. The second term dominates, so the numerator of Tajima's D asymptotes slowly to $\sim -nN_e\mu/\log n$, or generally a very negative number. The n and S dependence in the variance means that D ends up closer to 0, but values less than -2 are common.

This behavior, where most of the branch length on the tree is terminal, can also arise when rapidly expanding fit lineages overtake the population. A similar pattern applies to Fu and Li's D and F, because when almost the entire branch length of the tree is terminal, most mutations will be singletons whether selection or demographic expansion is responsible. Previous simulation studies have found that all of these statistics (except for H) depend sharply on assumptions about demography [105, 131].

In the literature, this confounding between demographic expansion and selection is so complete that positive selection is sometimes *equated* with a past reduced effective population size [7, 64]. The implication seems to be that selection and a reduction in diversity due to low population size are utterly indistinguishable, even in principle. As previously argued in Section 2.2, there are important differences between population expansion and selection that cannot be papered over simply by rescaling the population size.

The only real way to control for demography is to construct a scenario of demographic history, i.e., a schedule of changing population sizes, and to use these to determine merger rates in one's Kingman coalescent simulations. This is a risky endeavor, as it is very difficult to ascertain a population's demographic history. Moreover, real populations can undergo substantial demographic fluctuations, which can entail many parameters.

The only standard statistic that is nominally insensitive to demographic change is Fay and Wu's H. Recall that rapidly adapting populations are well described by the Bolthausen-Sznitman coalescent. In the BSC, a non-monotonic SFS results, as rapidly expanding fit lineages carry derived mutations that are then passed to all descendants: comparing the high and intermediate portions of the SFS means that H might in principle be sufficient for detecting selection. Population expansion and contraction cannot produce this pattern [52].

In fact, there is a close link between a non-monotonic SFS and an unusual tree topology. An excess of common derived variants is consistent with recent rapid expansion of a fit lineage, all of whose mutations will be inherited by its descendants. These descendants will likewise comprise a large portion of the population. As a general rule,

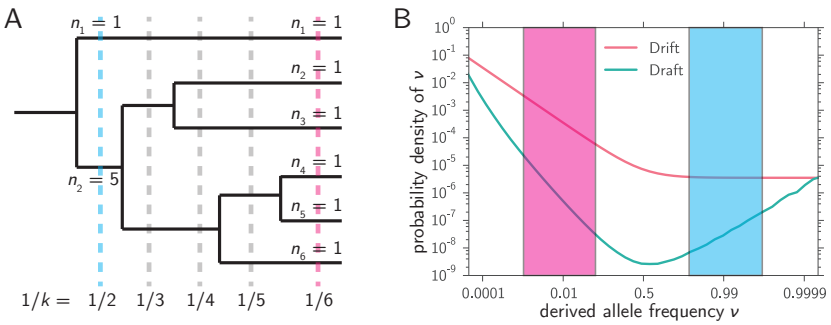

Figure 4.1: A schematic of the partition entropy. After the $k-1$ branching event, there are k lineages. We compute the sum of the Shannon entropies of the downstream branch populations, with a $1/k$ weighting factor (panel A). The $1/k$ term serves to favor early branching events, meaning that the partition entropy is, in a way, sensitive to the common end of the SFS (panel B). Note that the draft (BSC) curve in panel B differs slightly from Figure 2.4 in its normalization.

a changing population size distorts branch lengths on a genealogy, but it *cannot* distort a genealogy's topology [111].

Recall from Section 2.3 that the compound parameter $N\sigma$ is what determines whether neutral variation is governed primarily by drift or by draft, and hence whether genealogies are better described by the Kingman coalescent or the BSC [89] If $N\sigma \ll 1$, then the fittest individual in the population is unlikely to be much fitter than the average, so equal fitness is a reasonable approximation. If $N\sigma \gg 1$, the fittest individual may be far ahead of the distribution's bulk. The fitness position of the individual that is likely to give rise to the distribution's bulk (setting $\bar{x} = 0$ as before) is roughly $x = \sigma\sqrt{2\log N\sigma}$. Rapid turnover of the population is therefore to be expected, and most of the future population will descend from fit individuals near the nose. It stands to reason that it should be possible, based on a genealogy, to determine whether $N\sigma \ll 1$ or $N\sigma \gg 1$, or even to assign a value of $N\sigma$ to a genealogy. But classical imbalance statistics, which are sensitive to branch lengths and (with the exception of H) investigate deviations from neutrality primarily in the rare end of the spectrum, are unlikely to be particularly useful for this purpose, as they fail to distinguish between population expansion and rapid adaptation.

4.2 THE PARTITION ENTROPY

To this end, we define a novel statistic, the *partition entropy*, that depends only on the topology of a tree and therefore cannot be distorted away from a neutral expectation by demographic effects. "Topology" in this case refers not only to the branching pattern but also the order

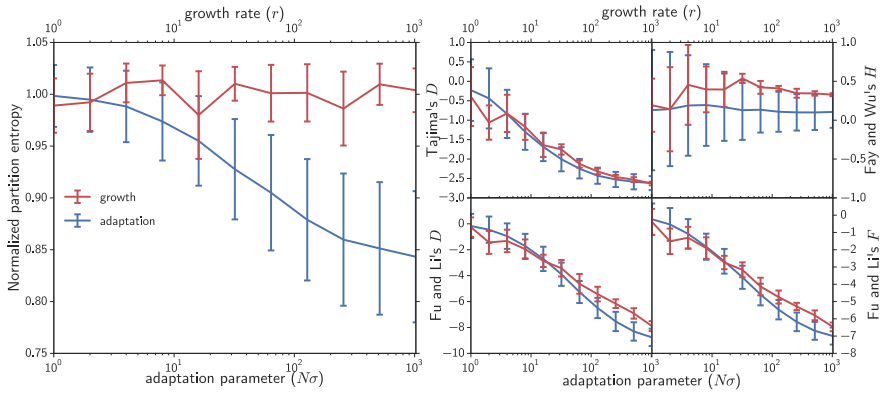

Figure 4.2: A comparison of the behavior of classical imbalance statistics and the partition entropy H (divided by its Kingman average) under selection. The classical statistics, for the most part, cannot distinguish at all between demographic expansion and adaptation (parametrized by $N\sigma$). The partition entropy can.

of branching events (i.e., the tree is ranked). There is no dependence on branch length whatsoever: it is therefore, by construction, independent of demography, as population expansion and contraction cannot distort the topology of a tree. We consider only slices of a tree ("partitions") in between branching events, then compute the Shannon entropy of the number of descendants of each extant branch.

If T is a binary, rooted, ultrametric tree with n leaves (and hence $n-1$ branching events), the *partition entropy* is

$$H(T) = \sum_{k=2}^{n} \frac{1}{k} \sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}, \tag{4.13}$$

where n_i is the number of observed individuals descended from branch i. The sum proceeds from the branching event just after the root, where there are $k = 2$ extant lineages, and scans forward until the present, where there are $k = n$ extant lineages (corresponding to the leaves of the tree). A schematic is provided in Figure 4.1.

The $1/k$ weighting term serves to favor earlier branching events, where strong imbalance events are most likely to be significant. Deviations from the Kingman coalescent in early partitions are likely to be due to expanding fit lineages, whereas deviations in late partitions a) are severely limited by construction (the range of possible values the branch offspring number can take on is small) and b) are likely to be statistical artifacts, i.e., due to drift or sampling noise. The partition entropy unfortunately has ugly moments that do not admit a simple representation, but they can be calculated explicitly, which we do in Appendix A. If an unbiased estimator is desired, one can simply subtract the expected value of the partition entropy from the result and compare its value to zero.

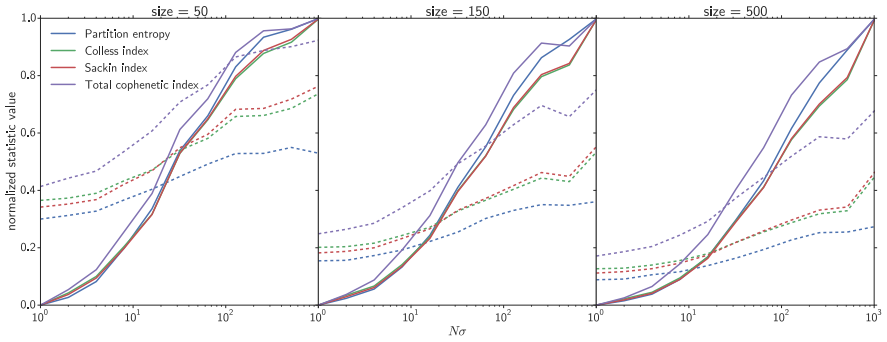

Figure 4.3: Comparison of the partition entropy and the Sackin, Colless, and cophenetic index for trees of varying size. Mean values are distorted so that they take on values in the interval $[0, 1]$: standard deviations (dotted lines) are divided by the same normalization. Though all have essentially the same dependence on the adaptation parameter $N\sigma$, the partition entropy consistently has the smallest standard deviation and therefore the best signal to noise ratio.

4.2.1 *Dependence on rates of growth and adaptation*

By construction, the partition entropy should not depend on demography at all: on the other hand, demography and adaptation should be very difficult to disentangle for branch length dependent statistics. We confirm this with simulations. To test the effects of demography, we relied on the program `ms` [62], simulating the Kingman coalescent with an exponential growth rate $r = \{0, 1, 2, 4, \ldots 1024\}$. To test the effects of selection, we relied on forward simulation with `FFPopSim` [128], with $N\sigma$ set to the same range of values and genealogies sampled at specified time intervals (see Section 4.2.3 for methods).

Our results confirm the claims set forth in Section 4.1.2. Most of the branch length dependent statistics have no ability at all to distinguish between demographic expansion and adaptation: the partition entropy does. Somewhat curiously, Fay and Wu's H turns out not to depend sharply either on $N\sigma$ or on r at all.

Next, we considered several other statistics that summarize the degree of imbalance in a tree, namely the Sackin, Colless, and total cophenetic indices [27, 87, 109]. The Sackin index is a sum over the number of nodes traversed between each leaf and the root of the tree: the Colless index is a sum of each node's "balance" (absolute value of the number of leaves downstream of each daughter node): and the total cophenetic index is the sum of the number of nodes traversed between the MRCA of each pair of leaves and the root. Note that the Colless and Sackin indices almost measure the same thing [87].

These are typically applied not to genealogies but to phylogenetic trees, for the purpose of distinguishing between models of speciation, in particular whether some lineages speciate more rapidly than oth-

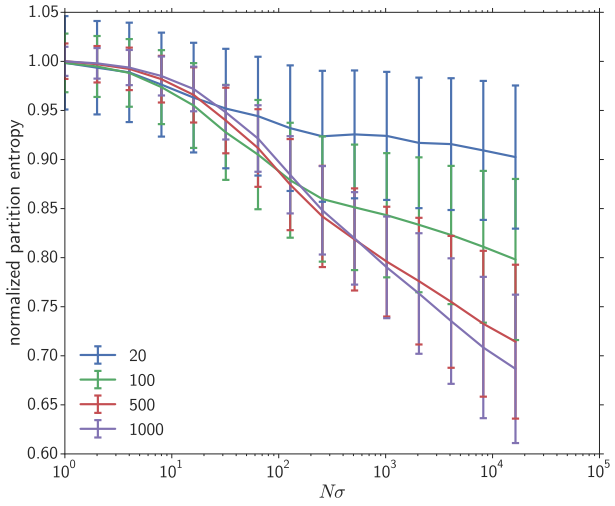

Figure 4.4: Partition entropy values for simulated trees of varying $N\sigma$ and size (see legend), divided by their Kingman mean. The relationship between them becomes log-linear starting at around $N\sigma = 10$; later, we argue that $\log(1 + N\sigma)$ is a better relation. Sampling more leaves effectively causes the dependence on $\log N\sigma$ to be linear over a greater range: more sampling is better.

ers. The simplest model of speciation is the Yule model, where each lineage gives rise to daughter lineages at a fixed rate [127]. This produces trees with statistics identical to the Kingman coalescent [132], with the exception that the time between branching events scales as the number of extant branches k rather than $k(k-1)/2$. There is no obvious *a priori* reason why they should not be applicable to tests of selection especially because, being dependent only on topology, they are likewise independent of assumptions about demography.

We test the imbalance statistics' and partition entropy's dependence on $N\sigma$ with forward simulations using `FFPopSim` [128]. A comparison of their dynamic range with the $N\sigma$ dependence of their standard deviations, as a rough signal to noise ratio, is visible in Figure 4.3. As it happens, while each of the three is sensitive to the kinds of distortions induced by selection, all of them are substantially noisier than the partition entropy. This sharply limits their utility for inferring the strength of selection. One reason for the partition entropy's superior performance is that the partition entropy relies on the order of branching events, which adds a minimal element of extra information to our analysis, and favors early branching events that are more likely indicative of the expansion of fit clones.

The partition entropy H and the adaptation parameter $N\sigma$ are well correlated. H appears to be roughly linear in $\log N\sigma$ (see Figure 4.4) or $\log(1 + N\sigma)$, and the correlation between them is much stronger at larger sample sizes n. In our comparison, we rely on the *real* genealogy of our trees rather than inferred genealogies. In real populations,

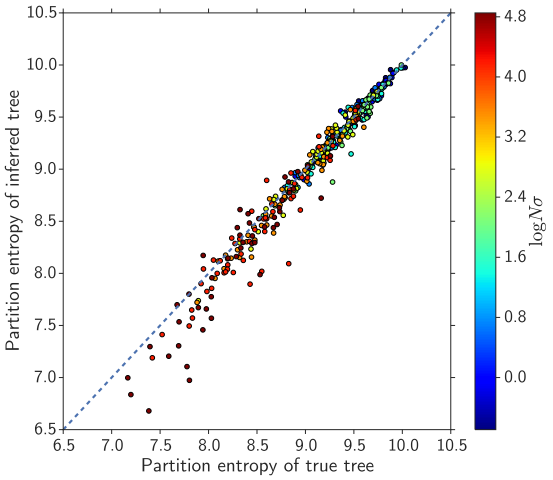

Figure 4.5: Partition entropy values for real and inferred genealogies and varying values of $N\sigma$: here the sample size is 100. Since we are not concerned with branch lengths, only the tree topology must be reliably inferred, so real and inferred trees agree quite well.

we will not have access to the true genealogy. However, since we only need the tree topology to compute H, the process of inferring a genealogy is unlikely to introduce substantial errors. We check this by using `FFPopSim` to generate both sequences and genealogies, inferring trees from the sequences via FastTree [102], and ultrametrizing them by tweaking their branch lengths.

Distortion of branch lengths is performed according to the following algorithm. Consider two adjacent terminal branches v_1 and v_2 with branch lengths l_{v_1} and l_{v_2}: "adjacent" means that they are sister nodes. We compute the average $(l_{v_1} + l_{v_2})/2$ and set both branch lengths equal to that. We then move one step up the tree and consider the node w_1 ancestral to v_1 and v_2, as well as its closest relative w_2. Let l_{w_1} be the sum of the branch lengths downstream of w_1, and likewise for w_2. Then we multiply the entire branch length downstream of w_1, including that of any daughter nodes, by $(l_{w_1} + l_{w_2})/2l_{w_1}$, and likewise for w_2. After this process, w_1 and w_2 are now equally distant from the present. We reiterate this process until the tree is ultrametric. This conservatively lengthens short branches and shortens long ones so that the overall topology of the tree is not changed and branching events occur in roughly the same order. As expected, inferred and true genealogies have partition entropy values that agree quite well with each other: see Figure 4.5.

4.2.2 *Statistical rejection of neutrality*

Standard SFS based statistics do not have analytically calculable null distributions. Typically, one either chooses a suitable approximate

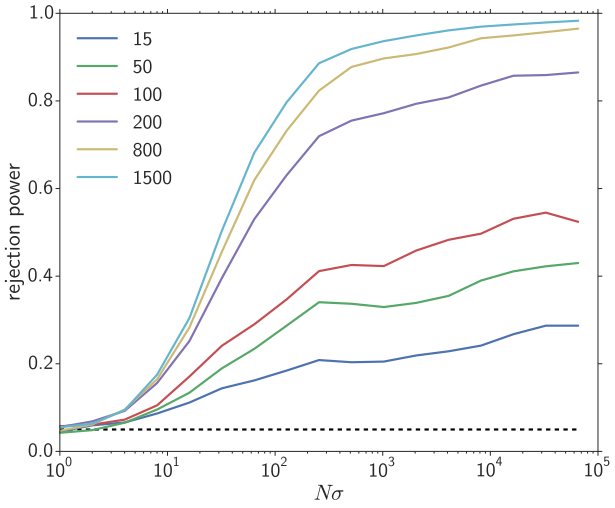

Figure 4.6: Power to reject the null hypothesis as a function of $N\sigma$, at the $\alpha = .05$ significance level, for various sample sizes (legend). We simulate trees of fixed $N\sigma$ and compare them to the Kingman partition entropy distribution to obtain the rejection probability. As $N\sigma$ and the number of leaves n increase, so does the power to reject the null (Kingman) hypothesis. Note that at very low $N\sigma$ or a small number of leaves, the probability is roughly equal to α (dotted line), as expected.

null distribution, such as a Gaussian or Beta distribution, or simulates the Kingman coalescent to generate a large number of samples, which can then be used to compile an "empirical" null distribution. The latter approach has gained traction in recent years because coalescent processes can be simulated very quickly.

Coalescent models furthermore can easily incorporate demography by rescaling the time between merger events. In general, let τ be the time elapsed between two merger events when the population size is fixed. If the true time between these events is t, during which the population size decreases from N_0 to N_t, then solving

$$\tau = \Lambda(t) - \Lambda(0) \tag{4.14}$$

yields the rescaled time t, where $\Lambda(t) = \int_0^t \frac{1}{\lambda(t')} dt'$ and $\lambda(t)$ is the time-dependent population size relative to the initial time point.

Unfortunately, demographic models sometimes have many free parameters, and inferring population size history is not straightforward, which mitigates the advantage afforded by incorporating demographic history into a model. On the other hand, the partition entropy is by construction independent of demography. So the Kingman coalescent with a constant merger rate, by itself, *is* a suitable null model for distinguishing between selection and neutrality. Hence a statistical test can be conducted in a manner similar to SFS based statistics, with no need to consider a demographic scenario at all.

We have implemented this in a simple Python script that takes as its input a tree, ultrametrizes it according to the procedure outlined in Section 4.2.3, and simulates the Kingman coalescent a large number of times to generate an empirical distribution. The neutral hypothesis can be rejected by considering the area under the curve more extreme than the computed value of H. The statistical power of this test increases strongly with the true $N\sigma$ associated with the tree, as illustrated in Figure 4.6.

4.2.3 *Simulation methods*

We determined the relationship between $N\sigma$ and the various imbalance statistics via forward simulation with `FFPopSim`. For the majority of our simulations, we considered a model where σ^2 is kept constant by an external constraint. We initialized populations with $N\sigma \in \left[2^{-4}, \ldots, 2^{17}\right]$ according to the following rule. For $N\sigma < 2^6$, we set $N = 4000$ and varied σ: for $N\sigma > 2^6$, we set $\sigma = 0.02$ and let N vary. All loci were assigned a random exponentially distributed fitness: all selection coefficients were adjusted every generation to keep the variance constant.

We employed an approximate infinite sites model where, every time a locus becomes monomorphic, a mutant at that locus is injected into a random individual in the population, and set each individual to have $L = 50000$ loci. In an asexual infinite sites model, the maximum possible number of populated clones is L, but this can decrease sharply due to selection rapidly driving clones to extinction or small population sizes. Decreasing the number of loci has little effect on our results except at very large sample sizes n, where the number of clones can be of order n. Genealogies were sampled every $\min(0.4N, 5/\sigma)$ generations, which we found empirically was about long enough for genealogies sampled from the same population no longer to be well correlated with each other. Populations were allowed to equilibrate a condition that we determined by checking whether the population's coalescence time was less than the number of generations elapsed; after the burn-in time, five sampling interval times elapsed before sampling began.

To test the effects of varying N and σ while keeping $N\sigma$ constant, we employed the same model but multiplied N by 2 and σ by 0.5 or vice versa. To verify that our results do not depend strongly on the choice of mutational model, we also performed simulations where mutations were injected into individuals at a constant rate $\mu = 1.0/L$ per site per generation, with $L = 3000$. All loci were initialized with negative selection coefficients; at exponentially distributed time intervals (with parameter $10/\sigma$), the selection coefficient of one locus was flipped.

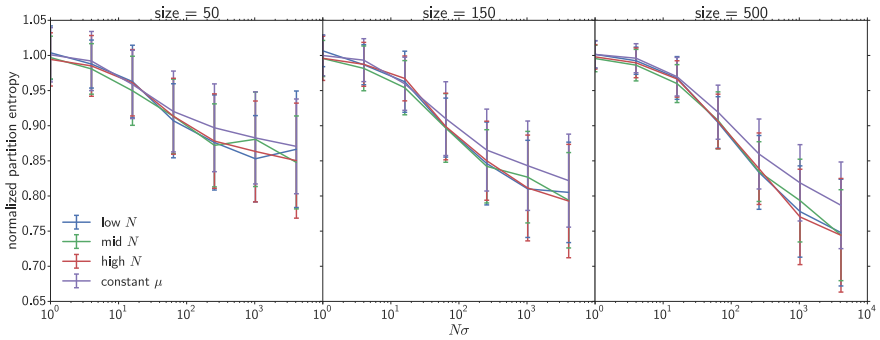

Figure 4.7: Dependence of the partition entropy on the mutational model and on the precise values of N and σ. The only difference noticeable is in the constant mutation rate model at high values of n, which is likely a byproduct of the fairly small number of loci we used.

For all simulations, a large number of genealogies (typically 1000) were generated by running several simulations in parallel. To avoid the need to repeat our analysis, we furnish a table of averages and standard deviations for varying sample sizes and values of $N\sigma$.

During simulation, it sometimes happens that more than one individual is sampled from the same clone. Because the relationship between individuals within a clone is not tracked, this appears on the genealogy as an unresolved "comb" node. This typically only happens in terminal or near terminal nodes. Any time it does, we conservatively replace the comb node with a random Yule-Kingman tree topology, with short branch lengths. This does not significantly distort the values of H we ultimately infer.

The precise values of N and σ turn out not to matter at all: only their product is meaningfully correlated with any of the imbalance statistics, as expected. Furthermore, the *ersatz* "infinite sites" model we employ is not a necessary assumption: the behavior is essentially the same even if mutations are injected at a constant rate, provided the fitness variance is kept constant: see Figure 4.7.

4.2.4 *Maximum likelihood of $N\sigma$*

For modest values of $N\sigma$ (up to roughly 10^3), H is roughly linear in the variable $\phi = \log(N\sigma + 1)$. This may be intuitive. The value of $N\sigma + 1$ determines whether drift (parametrized by $1/N$) or draft (parametrized by σ) is more important in shaping the fate of neutral alleles. This relationship breaks down somewhat at very high values of $N\sigma$, depending on the precise value of n: sampling more leaves extends the linear relationship. Refer back to Figure 4.4. The fact that H depends strongly on $N\sigma$ (via ϕ) means that it should be possible to use H to construct a maximum likelihood estimator $\hat{N\sigma}$ for a given tree, a possibility that we explore here.

To avoid the need to repeat our time intensive simulation process, we furnish a table of H mean and variance values for many combinations of $N\sigma$ and n. The dependence on n is not trivial, though, so the user must rely on splining between the mean and variance values we provide for nearby sample sizes. This provides a list of H mean and variance values for a new sample size.

We then leave the user one of two options. The user may either estimate $N\sigma$ by globally inferring a linear relationship between ϕ and H, with slope and intercept values estimated via maximum likelihood, and selecting the estimator $\hat{\phi}$ thereby. Or they may spline one step further, i.e., select $\hat{\phi}$ by assuming that there is a locally linear relationship between ϕ and H, depending only on the nearest values of H for the tree in question. This second approach is only viable if ϕ is monotonic in H, or else the relationship between them is not one to one. At low numbers of leaves, noise can remove this monotonicity, so the ability to infer $N\sigma$ is somewhat dependent on the sample size.

The former method is performed in a maximum likelihood framework. Suppose that errors in H are roughly Gaussian. Then the log of the likelihood ℓ becomes

$$\log \ell(D_n, m_n, b_n) = \sum_{i=1}^{l} \left(-\frac{(\mu_{ni} - m_n \phi_i - b_n)^2}{2\sigma_{ni}^2} \right) \tag{4.15}$$

(plus a constant), where D_n are the partition entropy values for sample size n, m_n and b_n are the slope and y-intercept, i ranges over the l values of $N\sigma$ used to generate training data, and μ_{ni} and σ_{ni} are the data mean and standard deviation of H. We minimize the log-likelihood to obtain estimates of m_n and b_n for a given sample size. All minimizations were performed using the Nelder-Mead algorithm implemented in `scipy` [96], with $(H_{\max} - H_{\min})/(N\sigma_{\max} - N\sigma_{\min})$ as our initial estimates.

No matter which method is selected, the subsequent process is the same. We seek the value of ϕ that maximizes the likelihood $\mathcal{L}(\phi|H) = P(H|\phi)$. Assume that the probability density $P(H|\phi)$ for a particular value of H is Gaussian, with linearly interpolated μ and σ values given by $\mu = mu_0 + \nu\phi$ and $\sigma = \sigma_0 + \tau\phi$. Then the maximum likelihood estimator $\hat{\phi}$ is achieved by solving

$$\begin{aligned} \hat{\phi}^2\tau^3 + \phi(2\tau^2\sigma_0 - \nu^2\sigma_0 - HT\nu + \tau\nu\mu_0) \\ +(\tau\sigma_0^2 + H\nu\sigma_0 + H^2\tau - \nu\sigma_0\mu_0 - 2H\tau\mu_0 + \tau\mu_0^2) = 0 \end{aligned} \tag{4.16}$$

for $\hat{\phi}$. Confidence intervals are obtained via $\hat{\phi} \pm c\sqrt{\mathcal{I}(\hat{\phi})}$, with the Fisher information

$$I(\hat{\phi}) = \frac{2\tau^2 + \nu^2}{\sigma^2} \tag{4.17}$$

and c a z-score critical value (≈ 1.96 for $p < 0.05$). The maximum likelihood estimator for $N\sigma$ is then given by $\hat{N\sigma} = \exp(\hat{\phi}) - 1$.

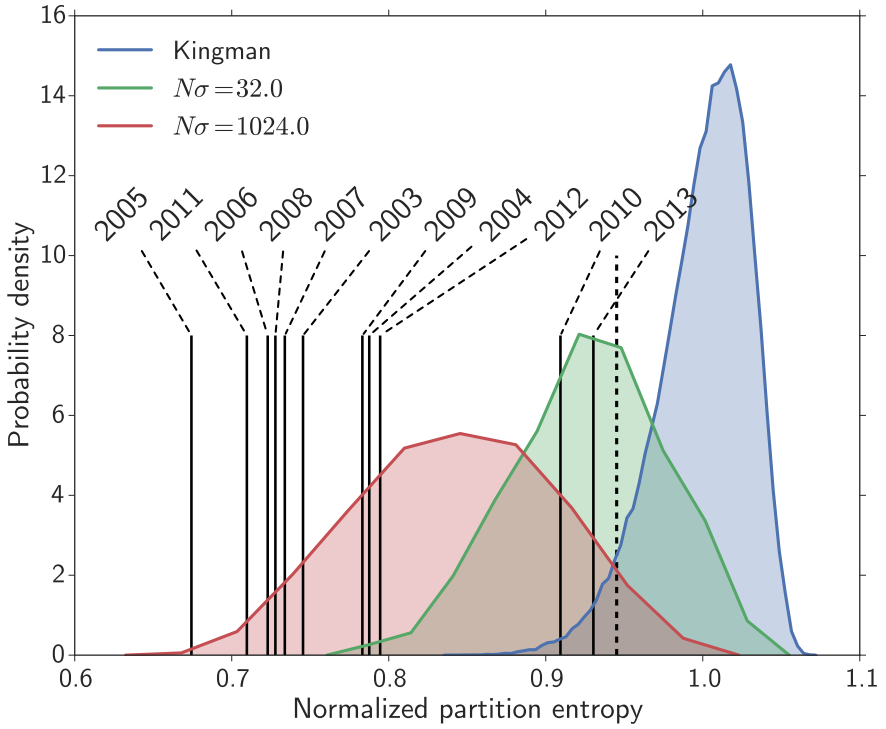

Figure 4.8: The distribution of the partition entropy under the Kingman coalescent and in rapidly adapting populations. The black dotted line indicates the cutoff for $\alpha = 0.05$ deviation from neutrality: high values of $N\sigma$ yield partition entropy values that make it very likely the null hypothesis will be rejected. The normalized partition entropy values for H3N2 sequences are indicated with black lines: all of them pass the $\alpha = 0.05$ threshold. We briefly discuss the shape of the distributions in Section 4.4.

4.3 RAPID ADAPTATION IN INFLUENZA A SUBTYPE H3N2

As a demonstration of our method, we investigated rapid adaptation in Asian influenza A sequences (subtype H3N2), segment HA. Influenza is an ideal candidate for our analysis. While flu viruses frequently reassort different segments of their genome when hosts are infected by multiple copies of the virus, individual segments evolve asexually [17, 118]. Additionally, the H3N2 subtype is clinically significant, as it is becoming increasingly common [12]. A combination of very high mutation rate and strong pressure to adapt suggest that $N\sigma$ may be quite far from 1.

We selected sequences from Asia, as pandemics tend to originate there [103, 108], and therefore we can be reasonably certain that we are sampling a well mixed population rather than distant offshoots. For each year from 2004 to 2013, we downloaded all H3N2 HA sequences from the Influenza Research Database [117], limiting our selected samples to one per geographical location per year (to avoid oversampling from a particular area and ensure some level of ran-

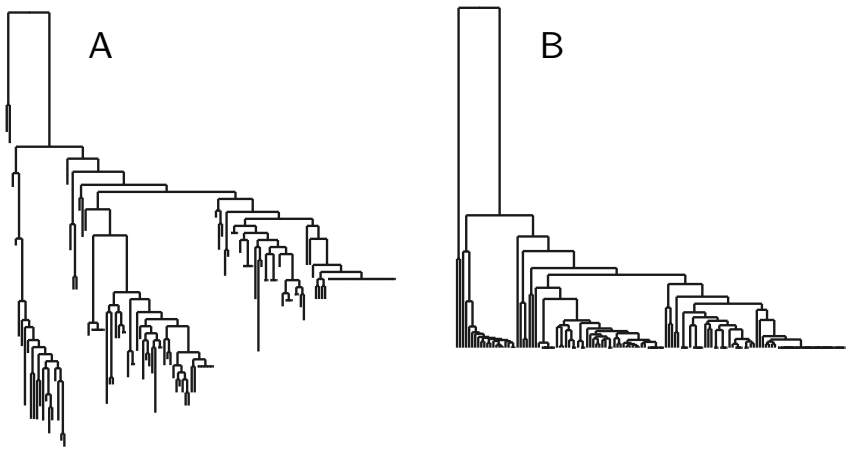

Figure 4.9: Inferred genealogies of H3N2 for sequences from 2010. Tree A is the original genealogy inferred from FastTree. Tree B shows the distortion of branch lengths we perform in order to guarantee ultrametricity. Either variation of the tree clearly indicates a strong deviation from neutrality.

domness). We aligned the sequences for each year using `MUSCLE` [35] and applied a rudimentary filtering criterion, excluding any sequences with more than three ambiguous nucleotides or gaps of larger than size three.

Next, we inferred the genealogy for a given year using `FastTree` [102], with default settings. We did not concern ourselves too much with the inference itself, as in principle only the topology is needed, which is easily inferred by neighbor joining. Any multiple mergers in the output genealogy (e.g., due to no mutations that differentiate an internal node) were replaced with a random Yule-Kingman tree topology, with branch lengths of approximately 10^{-6}. Then, the output genealogies were distorted to ensure ultrametricity: the decision to compare only sequences from within the same year was made to limit the effects of distorting the tree, as the coalescence time for flu is on the order of two to five years [10]. An H3N2 genealogy before and after applying the distortion algorithm (outlined in Section 4.2.1) is visible in Figure 4.9.

Based on the shape of these genealogies and our understanding of flu's dynamics, we straightforwardly predict that genealogies should overwhelmingly be consistent with very high values of $N\sigma$ and a rejection of genetic drift. To test the null hypothesis of genetic drift, we simulated 100000 Kingman coalescent trees for each sample size to generate a null distribution. We normalized the flu partition entropy values by dividing by the sample size dependent expectation and compared their values to a distribution of trees of size 100. The resulting p values are not surprising: in each case, we decisively reject the null hypothesis of genetic drift ($p \ll 10^{-5}$). A sketch of the test can be seen in Figure 4.8.

year	$N\sigma$	interval	sample size
2003	77948*	(7115, 853830*)	136
2004	2172	(455, 10350)	346
2005	204192*	(7081, 5887425*)	428
2006	27627	(2209, 345373*)	349
2007	18292	(1745, 191660*)	332
2008	176935*	(11930, 2623957*)	132
2009	2647	(500, 13995)	369
2010	82	(41, 182)	94
2011	46782	(3063, 714378*)	350
2012	9394	(2091, 43010)	129
2013	5	(0, 2051)	30

Table 4.1: Statistics of the inferred H3N2 genealogies. The inferred $N\sigma$ values and 95% confidence intervals are shown. Values marked with an asterisk are beyond the reliable range of our simulations ($N\sigma = 65536$), so we urge special caution with respect to them.

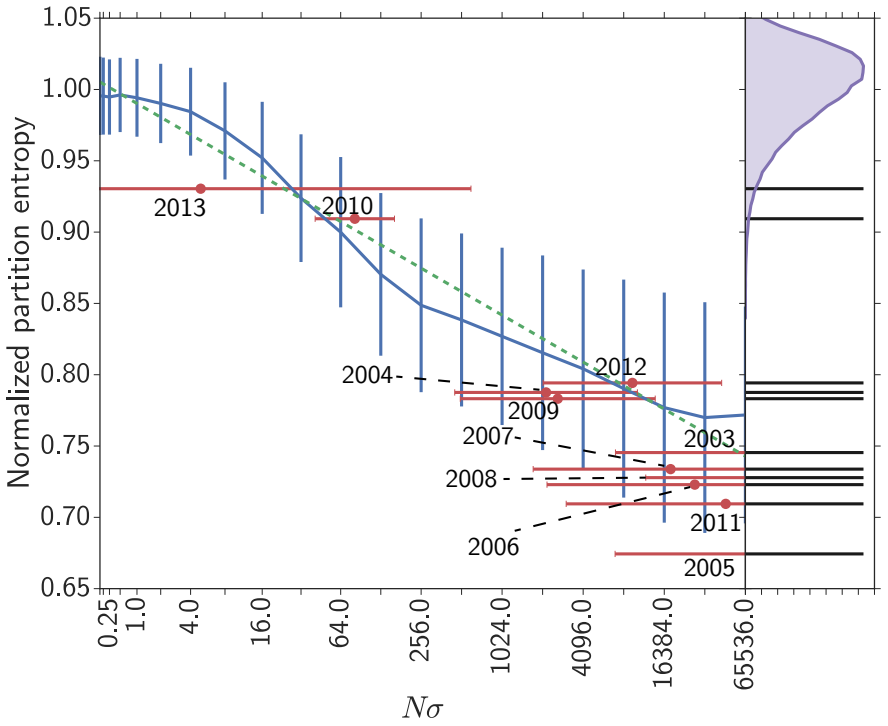

Figure 4.10: Relationship between the partition entropy and $N\sigma$ for trees of size 136. The green dotted line is the predicted relationship: red dots and bars denote the estimators $\hat{N\sigma}$ for each flu tree and confidence intervals. The corresponding black bars on the right show the position of each partition entropy value relative to the null (Kingman) distribution, similar to Figure 4.8.

These results immediately suggest that H3N2 trees might be more consistent with an extreme value of $N\sigma$ than with the Kingman coalescent. Determining which value of $N\sigma$ is most appropriate is our next task. We used our maximum likelihood model to assign an estimate for $N\sigma$ to each tree: results are shown in Figure 4.10 and are further summarized in Table 4.1.

Again unsurprisingly, we find very large values of $\hat{N\sigma}$, some of order 10^5: note that we only simulated $N\sigma$ up to 2^{17}, and several trees (from 2003, 2008, and 2005) appear to correspond to even larger $\hat{N\sigma}$ values. Note, furthermore, that even though we predict $N\sigma$ based on the inferred relationship between H and $N\sigma$ (green line), the inferred values (red dots) do not generally fall on this line. This is because the relationship generally depends on the sample size: the values of H shown are normalized relative to each tree's sample size, and the values of $N\sigma$ will depend on the simulation mean and variance values for that particular size. In general the confidence interval width is somewhat sensitive to the sample size: the confidence interval for the year 2013 actually hits $N\sigma = 0$.

We caution that the relatively wide variation in $N\sigma$ from year to year should not be taken to indicate that selection was many orders of magnitude more or less influential in different years; note the very broad confidence intervals. Rather, variation in $N\sigma$ is variation in the tendency for rapid expansion of clones, rather than slow diffusion-like drift, to determine the behavior of neutral alleles. We tackle the question of alternative interpretations of $N\sigma$ in the subsequent section.

4.4 CONCLUSION

We have described a novel, demography independent method for inferring the presence and strength of selection from the shape of a single genealogical tree. The demography independence is not accidental but rather by construction: demography cannot affect a tree's topology. Our emphasis on skewed branchings, especially early skewed branchings, makes our method sensitive to the shape of the SFS, primarily focusing on the common end of the spectrum.

The lack of understanding of the partition entropy's analytical behavior is a regrettable shortcoming of our work. In A, we demonstrate how to calculate its first and second moments under the Kingman coalescent, by considering the branch offspring distribution as a partition of n leaves into k categories: the partition, in the Kingman limit, is uniform on the simplex of integers $1, \ldots n$. Simplex representations for other coalescent processes such as the Bolthausen-Sznitman coalescent [113], are available, which may shed some light on the proper way to understand the partition entropy's analytics. A glance at Figure 4.8 suggests that different values of $N\sigma$ yield values of H that

might be roughly Beta distributed, with the interval given by the partition entropies for maximally balanced and maximally skewed trees.

Hitherto we have restricted our analysis solely to asexual populations. It may also be applied to asexual chromosomes in sexual populations. Nonetheless, most populations of interest undergo at least infrequent recombination. A further generalization is possible to well defined linkage blocks in sexual populations, which behave effectively asexually. This is a task we tackle in Chapter 5. We would, nonetheless, welcome extensions of this work to ancestral recombination graphs, where regrettably the theory is less robust but the general applicability is greater.

Naturally, caveats apply to our work in this section. We require that the genealogies under consideration be ultrametric, which somewhat mollifies the advantage afforded by relying solely on topology. However, many phylogenetic software packages easily facilitate the construction of ultrametric trees, e.g., `BEAST` [34]. We stress the need to consider only samples that are gathered on a time scale short relative to the total coalescence time of the population. An additional concern is sampling bias (also known as ascertainment bias), which is an issue for any method of inferring selection [94, 104]. Correcting for it, for example by incorporating an explicit model of how sampling occurs [19], is not straightforward. Nonetheless, through our emphasis on early branching events, we believe this issue is not likely to be significant for the partition entropy: we expect sampling bias to effect artefactual imbalance primarily in near-terminal nodes.

We have argued that the shape of flu genealogies is highly inconsistent with genetic drift. Clearly, alternative models of neutral variation are needed in order to understand their dynamics. One may immediately raise the question of whether the high values of $N\sigma$ we infer are indeed due to selection rather than some other process that might lead to rapid amplification of a lineage. In the case of flu, which lineage survives to dominate a population might not necessarily be a product of the ability to stay ahead of the host population's immune response but rather which individual is lucky enough to end up in virgin territory (e.g.., a naïve host population). We believe two factors mitigate this concern. First, we have selected flu samples in Asia, which close to the geographic origin of major outbreaks: flu lineages in this region constantly compete with each other, so random colonization events are less likely to determine which one dominates. Second, the evolution of influenza is canalized enough by selective constraints to be somewhat predictable [11]. These suggest that while stochastic processes, such as fortunate colonization of a fresh host population, have *some* effect on our high estimates of $N\sigma$, they are not entirely responsible.

Distinguishing between various non-diffusive population processes remains an interesting problem, however. Many biological systems,

such as Atlantic cod [3] or Pacific salmon [22], are well described by genealogies that do not appear Kingman. Broad offspring number distributions may be due to very large clutch sizes and quickly declining survivorship curves, and which individual's offspring dominate the future population is determined essentially by luck. This does not appear to be "selection", since their success is not heritable: the offspring number of an individual is not well correlated with that of its descendants.

One possibility for distinguishing between selection and rapid expansion due to some other process is to consider sexual populations and to compare the genealogy in different parts of the genome. Non-Kingman behavior throughout the genome might indicate strong selection and many sweeping loci throughout the entire genome, or it might indicate rapid expansion of lineages due to a more random process: but a difference from one part of the genome to the next may be an indicator of rapid adaptation in one portion and more neutral processes in the other.

This distinction is the kind of argument used by Batini et al. [7] and Karmin et al. [64], who noted that the human Y chromosome in Europe coalesces quickly to a small number of individuals: the genealogies they produce appears inconsistent the Kingman coalescent. This rapid coalescence is not observed in the mitochondria or autosomes, leading them to conclude that the sweeping of the Y chromosome is due to sex-specific selection factors. (They speculate that this is likely due to social and cultural factors, such as increased social stratification and new technology around the time these individuals lived: whether this really qualifies as "selection" is an interesting philosophical question.) As previously discussed in Section 2.3, their analysis equivocates between a depression in N_e and strong selection on the Y chromosome. Our work further heightens the need for researchers to differentiate these distinct processes.

5

RAPIDLY ADAPTING SEXUAL POPULATIONS

5.1 INTRODUCTION

In the previous chapter, I described a project that focuses on genealogies in asexual populations. Indeed, the very notion of a "genealogy" suggests that the population of interest is asexual: if individuals have two ancestors rather than one, something like an ancestral recombination graph is a more complete model of the population's evolutionary history. Here I describe a project to which I contributed primarily by performing simulations, the aim of which was to extend the theory of rapidly adapting asexual organisms to sexual populations.

In asexual populations, new beneficial mutations that arise compete with others and can slow or inhibit their sweeping, a phenomenon known as clonal interference [32, 51, 88]. In the limit of high recombination rate ρ, on the other hand, sweeping beneficial mutations essentially do not compete at all, and drift is the major force that shapes neutral variation. The intermediate picture is somewhat more complicated: distant loci decorrelate, but tightly linked ones have related histories. This interference between closely linked loci is also known as Hill-Robertson interference [61] and decreases the effectiveness of selection at linked parts of the genome.

The transition from drift to draft in asexual populations, including the effect on site frequency spectra and on the shape of genealogies, is heavily governed by $N\sigma$ [33, 89, 54]. Most populations are not truly asexual, however: even nominally "asexual" organisms tend to be facultatively sexual or undergo ersatz "outcrossing" through a related process like horizontal gene transfer. This is not surprising, as recombination affords many evolutionary advantages, including the avoidance of Hill-Robertson effects [41] (though see Schwander and Crespi [112] for an alternate perspective on the rarity of true asexuality). The ubiquity of sexual reproduction can be reconciled with the importance of genetic draft: if selection acts on many loci throughout the genome, then draft may be critical in shaping genetic variation even in sexual organisms. This may further help explain the "paradox of variation" [76], the lack of correlation between population size and genetic diversity, which runs directly counter to the neutral theory [68].

What follows is a simple scaling argument that reduces the problem of coalescence in sexual populations to the better understood process of coalescence in asexual populations. The central realization is that suitably defined portions of the genome behave effectively asex-

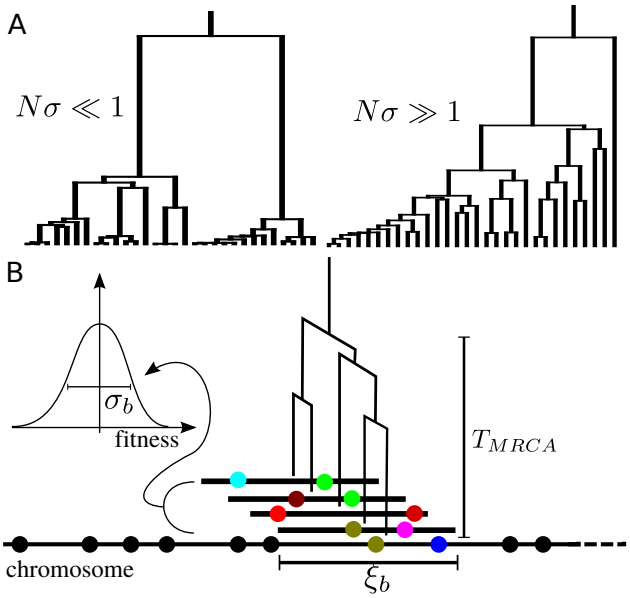

Figure 5.1: Coalescence and recombination. Panel A shows typical trees in the limits of drift and draft dominance, respectively: the $N\sigma \ll 1$ tree is roughly Kingman, and the $N\sigma \gg 1$ tree is roughly BSC. Panel B illustrates coalescence in sexual populations, with polymorphisms as colored balls. The genealogy changes as one slides along the chromosome. Over distances shorter than ξ_b, loci share most of their history: within this block length, neutral variation and coalescence are driven either by drift or by fitness differences in the genetic background. The size of ξ_b is related to the coalescence time T_{MRCA}: shorter coalescence times lead to larger blocks, since there has been less time for recombination to break up the block. It is also related to the proportion σ_b^2 of the total fitness variance σ^2 segregating in the block: a higher σ_b^2 corresponds to a more recent T_{MRCA}.

ually: "suitably defined" refers to a portion of the genome whose size depends on competition between recombination and selection. Thus, the behavior of sexual populations can be characterized in terms of asexual populations, with appropriately rescaled parameters. We test our predictions using forward simulations from `FFPopSim` [128] (see Section 5.3) which were my major personal contribution to this project.

5.2 COALESCENCE IN SEXUAL POPULATIONS

We outline the scaling argument briefly using heuristics before justifying it with slightly more rigor. In a sexual population, fit alleles tend to drag neutral variation with them, amplifying a nearby block of the genome. Recombination chops up these blocks, decorrelating the behavior of sufficiently distant loci (i.e., breaking up linkage). But over some block length, the rate at which a fit chunk of the genome

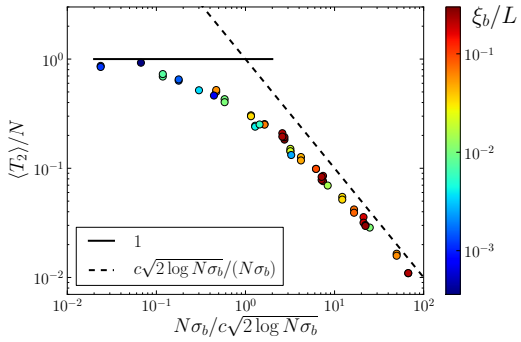

Figure 5.2: Pair coalescence times. As $N\sigma_b \to 0$, the block length ξ_b becomes shorter, and we recover the standard neutral coalescence time $\langle T_2 \rangle = N$. In the opposite limit, the coalescence time characteristic of rapid adaptation applies.

is amplified is roughly equal to the rate at which it is chopped up. Within this block length, most loci tend to share most of their genealogical history, and the block length likewise sets the scale over which linkage disequilibrium decays. Additionally, natural selection should shape the genealogy of this block. Rapid adaptation leads to rapid coalescence to fit individuals in the recent past: recombination, by breaking up associations between loci, recovers neutral dynamics and sends coalescence events further into the past. Such a block can be modeled as effectively asexual, meaning that many of the results of Neher and Hallatschek [89] should apply. This process is illustrated in Figure 5.1.

We proceed by identifying the appropriate block length based on the coalescence properties of an asexual population. Consider an asexual haploid population with fitness variance σ^2 due to many small effect mutations. If $N\sigma \gg 1$, the fittest individuals are the only ones whose lineages stand a good chance of surviving. In terms of fitness, they are roughly $x_c = \sigma\sqrt{2\log N\sigma}$ ahead of the mean [107, 121], and their lineages will take approximately $\sigma^{-1}\sqrt{2\log N\sigma}$ generations to dominate the population [89]. Thus, the probability that two random individuals had a common ancestor roughly $\sigma^{-1}\sqrt{2\log N\sigma}$ generations ago is of order 1. On the other hand, if $N\sigma \ll 1$, coalescence is dominated by genetic drift and takes approximately N generations. So the mean pair coalescence time is given by

$$\langle T_2 \rangle \approx \begin{cases} N & \text{if } N\sigma \ll 1, \\ c\sigma^{-1}\sqrt{2 log N\sigma} & \text{if } N\sigma \gg 1, \end{cases} \tag{5.1}$$

with c of order 1 depending on the details of the particular model. For the infinitesimal model we rely on, it is $\sqrt{12}$ [89].

These two limits determine the overarching behavior of the population's genetic variation, as well as its history. We will see that much

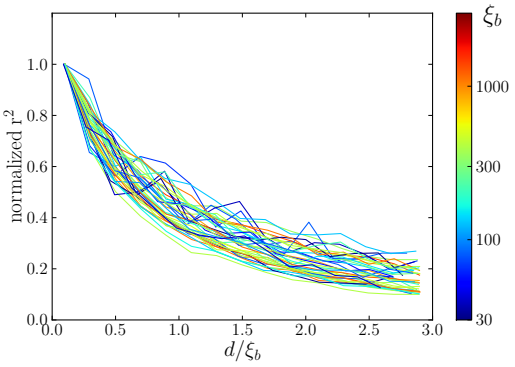

Figure 5.3: Linkage disequilibrium, measured by r^2 between loci as a function of the distance between them. The characteristic length was determined using Equation 5.2: after normalization with respect to ξ_b, all simulations follow essentially the same master curve, demonstrating that ξ_b sets the length scale for the decay of LD.

the same applies in a sexual population, with the exception that σ must be rescaled. Let ξ be the characteristic distance over which loci share most of their history. In general, ξ will decrease due to recombination:

$$\xi(t) = \frac{L}{1 + L\rho t} \approx \frac{1}{\rho t}. \tag{5.2}$$

If fitness variation is distributed rougly evenly throughout the population, a block of the genome will harbor a fraction σ^2_ξ of the total fitness variance. For a block of length ξ, this is

$$\sigma^2_\xi = \sigma^2 \frac{\xi}{L}. \tag{5.3}$$

The relevant σ_ξ is the amount of fitness variation σ_b that segregates in a block of length ξ_b that is unlikely to be broken up during the coalescence timescale.

If fitness variation in the block is substantial (i.e., $N\sigma_b \gg 1$, coalescence in this part of the genome will occur in $\langle T_2 \rangle = c\sqrt{2 \log N\sigma_b}/\sigma_b$ generations. Plugging $\langle T_2 \rangle$ into Equation 5.2 yields

$$\xi_b = \frac{\sigma_b}{c\rho\sqrt{2 \log N\sigma_b}}, \tag{5.4}$$

or equivalently

$$\sigma_b = \frac{\sigma^2}{L\rho c\sqrt{2 \log N\sigma_b}} \tag{5.5}$$

and

$$\xi_b = \frac{\sigma^2}{2L\rho^2 c \log N\sigma_b}. \tag{5.6}$$

If, on the other hand, coalescence in this block is *not* primarily due to exponential amplification of fit lineages but rather due to drift processes, then we recover $\xi_b \sim (N\rho)^{-1}$, as expected under genetic drift [60]. We confirm the behavior of $\langle T_2 \rangle$ with simulations: see Figure 5.2.

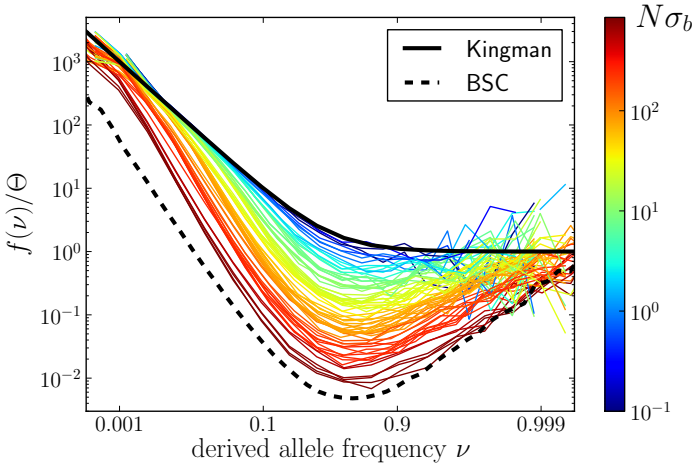

Figure 5.4: Site frequency spectra normalized by $\Theta = 2N\mu$ for many parameter combinations. Increasing $N\sigma_b$ causes spectra to converge to the BSC expectation. Note that the normalization of the BSC depends on $N\sigma_b$, so the BSC curve serves as a guide to the eye: compare with Figure 2.4.

In general, ξ sets the characteristic distance over which linkage disequilibrium in the genome decays exponentially. Within blocks of this length or less, most loci tend to share most of their genetic history, and LD remains elevated: see Figure 5.3. Sexual populations thereby behave quite similarly to asexual ones, with σ being replaced by σ_b. When $N\sigma_b < 1$, the Kingman limit applies, and when $N\sigma_b > 1$, the BSC arises in the genetic block under consideration. It is worth noting that this result obtains only in the limit $N\sigma > 1$. If this condition is not satisfied, then $\langle T_2 \rangle = N$, and linkage disequilibrium extends therefore over $l \sim (N\rho)^{-1}$ nucleotides. In effect, $N\sigma_b > 1$ is a more stringent constraint than $N\sigma > 1$: by decorrelating selected loci, recombination can cause genealogies to appear locally Kingman over a wide parameter range.

One way to appreciate the rescaling is to consider the site frequency spectrum as a function of $N\sigma_b$. We predict that the site frequency spectrum within an effectively asexual block should behave similarly to that of an asexual population, with $N\sigma_b$ determining whether the spectrum is closer to the Kingman coalescent or the BSC. Figure 5.4 illustrates this. As $N\sigma_b$ ramps up, the spectrum smoothly interpolates between the ν^{-1} behavior of the Kingman coalescent and the ν^{-2} or $1/\left[(\nu-1)\log(1-\nu)\right]$ scaling of the BSC.

Let us now consider the explicit dependence on the typical selection coefficient of a mutation s. If the variance σ_b^2 within a block is due to many small effect loci, then we expect [26, 121]

$$\sigma_b^2 \approx \frac{\xi_b \mu \left\langle s^2 \right\rangle}{2} \langle T_2 \rangle \,. \tag{5.7}$$

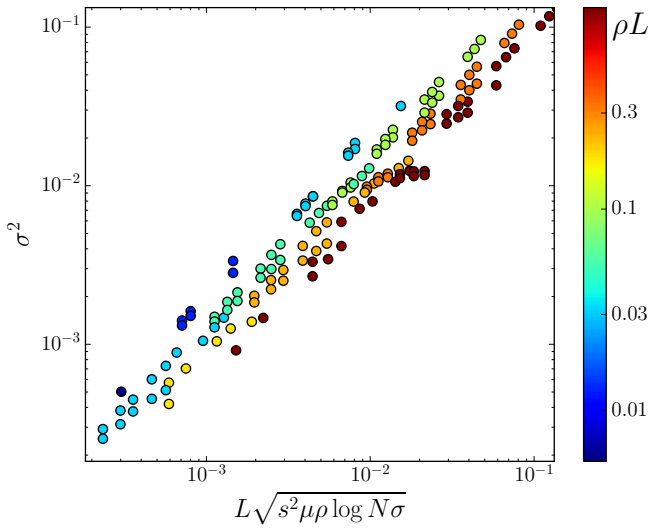

Figure 5.5: Fitness variation due to weak effect mutations in a "dynamic balance" model where beneficial and deleterious mutations keep the mean fitness roughly constant. The color shows the average number of crossovers per generation.

This applies whether s tends to be negative or positive, i.e., whether fitness variation is primarily due to many beneficial or deleterious mutations or a combination of both, since it depends only on the second moment $\langle s^2 \rangle$. Inserting Equation 5.2 into the above yields

$$\sigma_b^2 \approx \frac{\mu \langle s^2 \rangle}{2\rho}. \tag{5.8}$$

There is no dependence on the coalescence time at all, so the variance of an effectively asexual block is the ratio of the rate at which fitness variance is injected per nucleotide and the crossover rate. As a result,

$$\langle T_2 \rangle \approx \begin{cases} N & \text{if } N\sqrt{\mu \langle s^2 \rangle / \rho} \ll 1, \\ c\sqrt{\frac{\rho \log N\sigma_b}{\mu \langle s^2 \rangle}} & \text{if } N\sqrt{\mu \langle s^2 \rangle / \rho} \gg 1, \end{cases} \tag{5.9}$$

where c is again of order 1. The total rate of adaptation, when coalescence is driven by selection, is

$$\sigma^2 \approx cL\sqrt{\rho\mu \langle s^2 \rangle \log N\sigma_b}, \tag{5.10}$$

a result that holds whether mutations tend to be beneficial, deleterious, or a mix thereof: see Figure 5.5.

5.3 SIMULATION METHODS

To test our predictions, we relied on forward simulation. We considered several different models and implemented them in `FFPopSim`

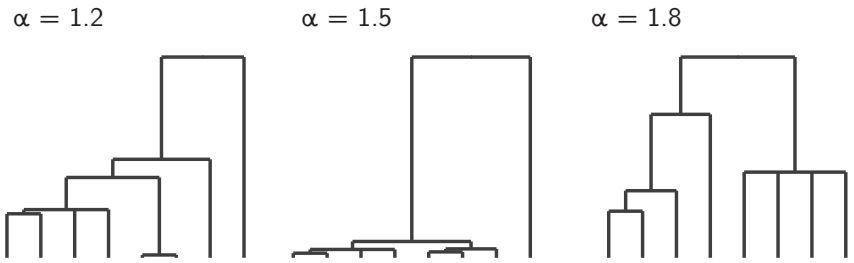

Figure 5.6: Beta coalescent trees for varying α, produced with `betatree`. As α moves from 1 to 2, branches become more balanced and multiple mergers occur less frequently.

[128]. For all models, we permitted populations to equilibrate for approximately $10T_{\text{MRCA}}$ generations before sampling: then, we sampled in intervals of roughly $\langle T_2 \rangle$ generations.

First, we considered a model where the injection of beneficial mutants keeps fitness variation σ^2 constant. As elaborated in Section 4.2.3, we force loci to remain polymorphic at all times by randomly introducing a mutation at a particular locus whenever that locus becomes monomorphic. The mutation rate is therefore kept small relative to $\langle T_2 \rangle$. We simulated a grid of parameters, with $N \in [1000, 3000, 10000]$, $\sigma \in [0.01, 0.03, 0.1]$, and $L\rho$ five logarithmically spaced values between 0.1σ and 1.0σ. Simulation results were filtered such that $\xi_b > 30$ and $\xi_b < L/3$.

Second, we considered a "dynamic balance" model where mutants are injected at a constant rate μ with small fitness effect s. The grid of parameters was $L \in [3000, 10000]$, $N \in [1000, 3000, 10000]$, $s \in [-0.001, -0.003, -0.01]$, $L\mu \in [1, 3, 10, 30]$, and $L\rho$ logarithmically spaced between s and 1.0. Simulations were filtered as in the first model, with the added constraint that $\langle T_2 \rangle \mu < 0.5$. We also considered a variant where mutations with positive s were injected (the other parameters were otherwise the same), as well as an infinite sites model where both deleterious and beneficial mutations were injected.

5.4 BETA COALESCENCE

The Kingman and Bolthausen-Sznitman coalescent describe the behavior of asexual genealogies when $N\sigma \ll 1$ and $N\sigma \gg 1$, respectively: in sexual populations, the limits $N\sigma_b \ll 1$ and $N\sigma_b \gg 1$ apply. For more moderate values of $N\sigma$ or $N\sigma_b$, an intermediate coalescent process may be a better model. In fact the Kingman and BSC are specific cases of the so called Beta coalescent, which in turn is a specific class of Λ-coalescents, in which

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\Lambda(dx) \tag{5.11}$$

and $\Lambda(dx)$ is the Beta$(2-\alpha, \alpha)$ distribution:

$$\Lambda(dx) = \frac{\Gamma(2)}{\Gamma(\alpha)\Gamma(2-\alpha)} x^{1-\alpha}(1-x)^{1-\alpha} dx. \tag{5.12}$$

Letting $\alpha = 2$ gives the Kingman coalescent; the BSC is the $\alpha = 1$ case. Lower values of α correspond to more extreme merger statistics, culminating in the *star coalescent* at $\alpha = 0$, where all lineages merge at once. Intermediate values result in trees that are somewhere between the BSC and Kingman extremes: see the examples in Figure 5.6.

As a component of this project, we wrote `betatree`, a Python package that allows the user to quickly simulate many Beta coalescent trees for a selected value of α. Trees can easily be used to generate site frequency spectra, which can be used to compute summary statistics for comparison with null models. Many of the site frequency spectra and coalescent trees in this thesis were generated using `betatree`.

5.5 CONCLUSION

We have provided a robust but simple scaling argument for understanding the process of coalescence in adapting sexual populations. Provided the condition $N\sigma \gg 1$ is satisfied, results from asexual populations apply to well defined linkage blocks: the parameter $N\sigma_b$ is what sets the time scale of coalescence, the speed of adaptation, the shape of the SFS, and the length scale over which LD decays. In conjunction with the results outlined in Chapter 4, we can move a step further and attempt to determine the value of $N\sigma_b$ in an effectively asexual block. In a roundabout way, we can even use Equation 5.9 to probe the distribution of mutational effects in a block, which might allow for comparison with the results of Rice et al. [106].

In our model, fitness variation is not due to a handful of major loci that sweep strongly against a neutral backdrop but rather a large number of polymorphisms. This is in contrast to other models of adaptation in sexual organisms such as Weissman and Barton [123], which have typically focused on a few mutations of large effect. Our results are similar in the limit of fairly small linkage blocks that are occasionally perturbed by selective sweeps.

Previously we have argued that genetic draft due to the effects of many weak loci may help explain the "paradox of variation", the lack of correlation between genetic diversity and population size [76]. This also has implications for quantitative genetics (from whence the "infinitesimal model" originates). If many loci in the genome experience weak selection, they may be cryptically correlated with quantitative traits. In effect, our work implies that genomic "dark matter" of unknown function might be a major part of the solution to the "missing heritability" problem [36, 83, 126] and may help to erect a bridge between quantitative genetics and the population genetics of rapid adaptation.

One interesting consequence of our work concerns the effects of nearly neutral loci. Recombination is infrequent in many organisms of interest: in others, it varies by several orders of magnitude across the genome [28]. As a result, haplotypes or well defined linkage blocks persist for long periods of time across regions of low recombination. This means that the block length ξ_b can be quite large. The density of segregating sites in the genome scales roughly as $2\mu \langle T_2 \rangle$, so the number of segregating sites in the block length $\xi_b \approx \rho / \langle T^2 \rangle$ is roughly $n = \mu/\rho$. The fitness variance in the block is given by $\sigma_b^2 = \langle s^2 \rangle n$, so the condition $N\sigma_b > 1$ implies $N^2 \langle s^2 \rangle n > 1$. This means that, somewhat surprisingly, selection can dominate neutral variation even if it is due to (nominally) nearly neutral mutations, for which $s < 1/N$: their combined effects cause individuals to have substantial fitness variance. Note that the second moment $\langle s^2 \rangle$ of mutational fitness effects does not depend on the sign of s: it applies whether mutations tend to be beneficial or deleterious.

It has long been argued that nearly neutral mutations present a stumbling block for the evolutionary process. As deleterious mutations of small effect fix (a variant of Muller's ratchet), the fitness in a population is slowly reduced: and indeed, in the limits of many loci, high mutation rate, and low population size, deleterious mutations should fix quite easily. This is the parameter regime that characterizes many large eukaryotes. Truncation selection, a form of synergistic epistasis (also referred to as "Kondrashov's hatchet"), is one purported mechanism that keeps fitness from decaying irreversibly [70]; occasional large effect beneficial mutations are another [56]. Our results suggest an additional possibility. Deleterious mutations confined to some linkage block may actually inhibit fixation because the total number of deleterious alleles in the block can vary substantially: even though individual deleterious mutations have effects too small to be noticed by selection, variation in mutation number can lead to significant fitness variation, meaning that selection can counteract the effects of drift. A large block harbors more variation, suggesting that, while a *high* recombination rate aids beneficial mutations by inhibiting Hill-Robertson effects, a *low* recombination rate may be useful for counteracting nearly neutral deleterious mutations.

One nagging question to which we have devoted considerable time is the relationship between Beta coalescent trees and $N\sigma$ (or $N\sigma_b$). It stands to reason that if the Kingman coalescent and BSC represent extremes in $N\sigma$, more moderate values of $N\sigma$ should correspond to an intermediate coalescent with a value of α between 1 and 2. This would in principle not be difficult to show with simulations, given the right choice of summary statistic: in fact this is similar to what Árnason and Halldórsdóttir [3] attempted to do. But the intermediate Beta coalescents have analytical properties that are not as well behaved as their extreme variants are, which limits our ability to make direct

analogies between them and real (or simulated) populations. Still, we would gladly welcome additional work in this area.

6

OUTLOOK

Classic neutral theory is a convenient framework for interpreting sequence data. Genetic drift is a well understood process, and it can be relatively straightforward to identify sweeping beneficial mutations against a neutral background. Nonetheless, evidence continues to mount that this characterization of the evolutionary process is not correct. This evidence is particularly strong for large, rapidly adapting populations such as many pathogens, where selection at linked sites pervades the genome, but other organisms may likewise be well served by a better understanding of genetic draft.

The main results of this doctoral work are several fold. First, we have established how traditional methods for understanding the fierce competition between the HIV virus and the host immune system can be misguided. A simple picture of the virus' evolution, where escape mutations sweep freely through the intra-patient population, engenders confusion about the immune response during acute infection. Our realistic but still tractable model for inferring the strength of viral escape suggests that the arms race between host and virus remains intense well into acute infection. We hope that similar approaches will see light in the study of other pathogens.

Second, we have provided a reasonable and parsimonious method for determining whether a particular pattern of neutral diversity is more consistent with genetic drift or genetic draft. In doing so, we have partially answered the question: to what extent can genealogies themselves inform us about the evolutionary process they represent? Approaches similar to ours have gained some traction in recent years, as it is increasingly realized that genealogical imbalance is informative in a way that traditional methods for assaying the presence of selection may not be. Our work comes at a fortuitous time, as the tendency to treat selection merely as a change in the "effective population size" can otherwise cause adaptation in the genome to be overlooked. Extensions of this work may be useful in the study of sexual organisms, where individual segments of the genome have their own genealogies, may further close the gap between rapidly adapting asexual organisms and bulkier species.

This gap is bridged partially by our third major result. The coalescence properties of asexual organisms have been remarkably well developed in preceding decades. We have successfully demonstrated that these properties can be extended to sexual organisms in the right limits, subject to a suitable rescaling of parameters. Our work also hints at a way to identify important population genetic parameters,

such as $N\sigma$, in sexual populations. Further development might focus on ancestral recombination graphs, the distinction between different methods of recombination, more complex genetic maps (e.g., ones that involve hotspots), or specific forms of fitness variation.

It is clear that genetic draft plays a critical role on some branches of the tree of life. But it remains difficult to say just how pervasive it is or what broader implications it may have for evolutionary theory as a whole. Is genetic draft important primarily in large microbial populations, or does it have major effects on other organisms, perhaps even humans? Cries of "the neutral theory is dead: long live the neutral theory" have been issued for decades now [72], and they are unlikely to stop any time soon. I began my doctoral work with the hope of furthering a new kind of population genetics, something more complicated but more robust than the standard neutral theory. Now, I hope to continue doing so.

A

APPENDIX: MOMENTS OF THE PARTITION ENTROPY

We proceed to derive explicitly the first and second moments of $H(T)$. With some modification, these results can be generalized to any linear sum over the partitions in a Yule-Kingman tree: they can, for example, be used to verify the ν^{-1} scaling law of the site frequency spectrum. Additionally, with a different partitioning of the simplex, they can be extended to other coalescent processes.

EXPECTED VALUE OF THE PARTITION ENTROPY

By definition, the expected value

$$\mathrm{E}(H(T)) = \sum_{T \in \mathcal{T}_n} P(T)H(T), \tag{A.1}$$

in which $\mathcal{T}_n$ is the set of all binary rooted trees with n leaves. Note that a particular tree T can be decomposed into a set of partitions $\pi_{k,n}$ of n leaves into k groups:

$$\sum_{T \in \mathcal{T}_n} P(T)H(T) = \sum_{\Pi_{2,n},\ldots\Pi_{n,n}} P(\pi_{2,n},\ldots\pi_{n,n})H(\pi_{2,n},\ldots\pi_{n,n}), \tag{A.2}$$

where $\Pi_{k,n}$ is the set of all partitions $\pi_{k,n}$. Each partition $\pi_{k,n}$ depends only on its predecessor $\pi_{k-1,n}$, so the expected value becomes

$$\sum_{\Pi_{2,n}} P(\pi_{2,n}) \sum_{\Pi_{3,n}} P(\pi_{3,n}|\pi_{2,n}) \ldots \sum_{\Pi_{n,n}} P(\pi_{n,n}|\pi_{n-1,n}) \sum_{k=2}^{n} H(\pi_{k,n}), \tag{A.3}$$

where the sums are nested, not multiplied. In general, for any k the sum involving $P(\pi_{k,n}|\pi_{k-1,n})$ will be multipled by $H(\pi_{k,n})$ but not any further terms in the internal sum, With some rearranging of terms, we therefore have

$$\sum_{T \in \mathcal{T}_n} P(T)H(T) = \sum_{k=2}^{n} \sum_{\Pi_{k,n}} P(\pi_{k,n})H(\pi_{k,n}). \tag{A.4}$$

Unfortunately the multiplicity of $\Pi_{k,n}$ can be very large, but we can distill this somewhat. Consider an arbitrary branch in the kth partition of the tree. There are k such branches, and in the Yule-Kingman scenario, an arbitrary branch has size m with probability

$$P_{k,n}(m) = \frac{\binom{n-m-1}{k-2}}{\binom{n-1}{k-1}} \tag{A.5}$$

for $1 \leq m \leq n-k+1$. For notation's sake, it is useful to define

$$h_{k,n}(m) = -\frac{1}{k}\frac{m}{n}\log\frac{m}{n}, \tag{A.6}$$

the component of the partition entropy due to a branch containing m individuals in the kth partition of n leaves. Thus,

$$\sum_{\Pi_{k,n}} P(\pi_{k,n})H(\pi_{k,n}) = k\sum_{m=1}^{n-k+1} P_{k,n}(m)h_{k,n}(m), \tag{A.7}$$

and the expected value becomes

$$\mathrm{E}(H(T)) = \sum_{k=2}^{n} k\sum_{m=1}^{n-k+1} P_{k,n}(m)h_{k,n}(m). \tag{A.8}$$

This admits no simpler representation, but it can at any rate be computed in $\mathcal{O}(n^2)$ time.

VARIANCE OF THE PARTITION ENTROPY

The variance

$$\mathrm{Var}(T) = \mathrm{E}(H^2(T)) - \mathrm{E}(H(T))^2. \tag{A.9}$$

The first term can be broken up into parts:

$$\mathrm{E}(H^2(T)) = \sum_{\Pi_{2,n},\ldots\Pi_{n,n}} P(\pi_{2,n}\ldots\pi_{n,n})H^2(\pi_{2,n}\ldots\pi_{n,n}). \tag{A.10}$$

Expanding out H^2 would reveal

$$\begin{aligned}\mathrm{E}(H^2(\pi_2,\ldots\pi_n)) = \mathrm{E}&\left(\sum_{k=2}^{n} H^2(\pi_{k,n})\right)\\ &+ 2\mathrm{E}\left(\sum_{k=2}^{n}\sum_{j=k+1}^{n} H(\pi_{k,n})H(\pi_{j,n})\right).\end{aligned} \tag{A.11}$$

In the first term, $H^2(\pi_{k,n}) = \sum_{x\in\pi_{k,n}}\sum_{x\in\pi_{k,n}} h_{k,n}(m_x)$: we must sum over the branches x in the k partition, which have m_x downstream individuals. This will yield k "diagonal" terms (where the same branch is sampled twice), and we know the probability distribution for the values of m. Hence there will be a contribution $k\sum_{m=1}^{n-k+1} P_{k,n}(m)h_{k,n}^2(m)$ to the sum. There are also $k(k-1)$ "off-diagonal" terms, where two different branches are sampled: and in general, if $P_{k,n}(m_1)$ is the probability distribution for size of the first branch, the second becomes $P_{k-1,n-m_1}(m_2)$, as there are now $n-m_1$ individuals to distribute into $k-1$ groups. Hence the off-diagonal contribution is $k(k-1)\sum_{m_1=1}^{n-k+1}\sum_{m_2=1}^{n-k-m_1+2} P_{k,n}(m_1)P_{k-1,n-m_1}(m_2)h_{k,n}(m_1)h_{k,n}(m_2)$. (Note that the dependence of m_2 on m_1 shows up in the probability term but not in h itself.)

We therefore obtain

$$\mathrm{E}\left(\sum_{k=2}^{n} H^2(\pi_{k,n})\right) = \sum_{k=2}^{n} k \sum_{m=1}^{n-k+1} P_{k,n}(m) h_{k,n}^2(m) + k(k-1) \sum_{m_1=1}^{n-k+1} \sum_{m_2=1}^{n-k-m_1+2} P_{k,n}(m_1) P_{k-1,n-m_1}(m_2) h_{k,n}(m_1) h_{k,n}(m_2). \tag{A.12}$$

The cross term $2\mathrm{E}(\sum_{k=2}^{n} \sum_{j=k+1}^{n} H(\pi_{k,n}) H(\pi_{j,n}))$ will be much more complicated, as $\pi_{k,n}$ and $\pi_{j,n}$ within a particular tree are not independent. We can contract the expectation with no problems. In general,

$$\mathrm{E}(H(\pi_{k,n})H(\pi_{j,n})) = \sum_{\Pi_{k,n}} P(\pi_{k,n}) H(\pi_{k,n}) \sum_{\Pi_{j,n}} P(\pi_{j,n}|\pi_{k,n}) H(\pi_{j,n}). \tag{A.13}$$

This can be rewritten as

$$\mathrm{E}(H(\pi_{k,n})H(\pi_{j,n})) = \mathrm{E}(H(\pi_{k,n})\mathrm{E}(H(\pi_{j,n})|\pi_{k,n})). \tag{A.14}$$

As before, we can decompose this into the possible branch sizes at the k and j partition. The internal expectation will simply be a sum over the possible sizes m_y that a branch y in $\pi_{j,n}$ can take on, conditioned on the size m_x of a branch x in $\pi_{k,n}$.

There are plainly two possibilities: either y is descended from x or it is not. To compute the distribution $P(m_y|m_x)$ in the case of descent from x, we need to know how many branching events l have arisen from x at partition k: then the distribution becomes $P_{l+1,m_x}(m_y)$, as x has $l+1$ descendants. Likewise, in the case that y is not descended from x but from one of the $k-1$ other branches at partition k, the distribution becomes $P_{j-l-1,n-m_x}(m_y)$, as there are $n-m_x$ remaining individuals and $j-l-1$ branches that do not descend from x.

We need the probability mass $P(l)$ of branch sizes l. Clearly, if a branch x at partition k has m_x offspring, the probability that it will be the next lineage to branch is $\frac{m_x-1}{n-k}$. Some induction reveals that

$$P(l|m_x,n,k,j) = \binom{j-k}{l} \times \frac{(m_x-1)!}{(m_x-1-l)!} \frac{(n-k-m_x+1)!}{(n-j-m_x+l+1)!} \frac{(n-j-1)!}{(n-k)!}, \tag{A.15}$$

provided all the arguments are 0 or higher. Iterating over the possible values of l, then the possible values of m_y conditioned on descent or lack of descent from x, yields the cross term:

$$\begin{aligned}
&\mathrm{E}(H(\pi_{k,n})\mathrm{E}(H(\pi_{j,n})|\pi_{k,n})) = \\
&k \sum_{m_x=1}^{n-k+1} P_{k,n}(m_x) h_{k,n}(m_x) \sum_{l=0}^{j-k} P(l|m_x, n, k, j) \\
&\times \left((l+1) \sum_{m_y=1}^{m_x-l+1} P_{l+1,m_x}(m_y) h_{k,n}(m_y) \right. \\
&\left. + (j-l-1) \sum_{m_y=1}^{n-m_x-j+l-1} P_{j-l+1,n-m_x}(m_y) h_{k,n}(m_y) \right).
\end{aligned} \tag{A.16}$$

With the cross term available, the first term of the variance

$$\mathrm{E}(H^2(T)) = \sum_{k=2}^{n} \left(\mathrm{E}(H^2(\pi_{k,n})) + 2 \sum_{j=k+1}^{n} \mathrm{E}(H(\pi_{k,n})H(\pi_{j,n})) \right) \tag{A.17}$$

can be obtained. Sadly, the iteration over $P(l)$ substantially worsens the run time, which appears to be $\mathcal{O}(n^4)$.

A simple argument can help make these scalings tangible. At a particular partition k, there are k individuals, each of which has roughly n/k offspring and contributes roughly $k^{-1} \log k$ to the total H. But fluctuations are on the same order of magnitude. Thus, the variance from the k partition is roughly $k^{-1} \log^2 k$, which integrates $\sim \log^3 n$. This suggests that the ratio of the mean and standard deviation decays slowly, as roughly $(\log n)^{-1/2}$: compare with *Figure* 4.4.

PUBLICATIONS

Some ideas and figures have previously appeared in the following publications:

- *Kessinger T*, Perelson A, and Neher R. Inferring HIV escape rates from multi-locus genotype data. *Frontiers in Immunology* 4:252, 2013.
  - Text and figures from this manuscript appear in Chapter 3.
- Neher R, *Kessinger T*, and Shraiman B. Coalescence and genetic diversity in sexual populations under selection. *Proceedings of the National Academy of Science*, 110(39):15836-15841, 2013.
  - Text and figures from this manuscript appear in Chapter 5.

In addition, the following manuscript will shortly be submitted:

- *Kessinger T*, Neher R. Genetic draft and the shape of genealogies. 2015
  - Text and figures from this manuscript appear in Chapter 4.

BIBLIOGRAPHY

[1] M. Altfeld et al. Enhanced detection of human immunodeficiency virus type 1-specific t-cell responses to highly variable regions by using peptides based on autologous virus sequences. *Journal of Virology*, 77(13):7330–7340, 2003.

[2] Aristotle. *Phusika akroasis*. 350 B.C.E.

[3] E. Árnason and K. Halldórsdóttir. *Nucleotide variation and balancing selection at the Ckma gene in Atlantic cod: analysis with multiple merger coalescent models.*, volume 3:e786. PeerJ, 2015.

[4] B. Asquith and A. R. McLean. In vivo CD8+ T cell control of immunodeficiency virus infection in humans and macaques. *Proc Natl Acad Sci USA*, 104(15):6365–6370, 2007.

[5] B. Asquith, C. T. T. Edwards, M. Lipsitch, and A. R. McLean. Inefficient cytotoxic T lymphocyte-mediated killing of HIV-1-infected cells in vivo. *PLoS Biol*, 4(4):e90, 2006.

[6] N. H. Barton. Linkage and the limits to natural selection. *Genetics*, 140(2):821–41, 1995.

[7] C. Batini et al. Large-scale recent expansion of European patrilineages shown by population resequencing. *Nat Commun*, 6, 2015.

[8] R. Batorsky et al. Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. *Proc Natl Acad Sci USA*, 108(14):5661–6, 2011.

[9] D. M. Beazley. Swig: An easy to use tool for integrating scripting languages with c and c++. In *Proceedings of the 4th Conference on USENIX Tcl/Tk Workshop, 1996 - Volume 4*, TCLTK'96, pages 15–15, Berkeley, CA, USA, 1996. USENIX Association.

[10] T. Bedford, S. Cobey, and M. Pascual. Strength and tempo of selection revealed in viral gene genealogies. *BMC Evol Biol*, 11: 220, 2011.

[11] T. Bedford, A. Rambaut, and M. Pascual. Canalization of the evolutionary trajectory of the human influenza virus. *BMC Biol*, 10:38, 2012.

[12] T. Bedford et al. Integrating influenza antigenic dynamics with molecular evolution. *eLife*, 3, 2014.

[13] E. Blyth. An attempt to classify the "varieties" of animals, with observations on the marked seasonal and other changes which naturally take place in various British species, and which do not constitute varieties. *Magazine of Natural History*, 8(1):40–53, 1835.

[14] E. Bolthausen and A.-S. Sznitman. On Ruelle's probability cascades and an abstract cavity method. *Communications in Mathematical Physics*, 197(2):247–276, 1998.

[15] V. F. Boltz et al. Ultrasensitive allele-specific PCR reveals rare preexisting drug-resistant variants and a large replicating virus population in macaques infected with a simian immunodeficiency virus containing human immunodeficiency virus reverse transcriptase. *J Virol*, 86(23):12525–12530, 2012.

[16] S. Bonhoeffer, A. D. Barbour, and R. J. D. Boer. Procedures for reliable estimation of viral fitness from time-series data. *Proc Biol Sci*, 269(1503):1887–93, 2002.

[17] M. F. Boni, Y. Zhou, J. K. Taubenberger, and E. C. Holmes. Homologous recombination is very rare or absent in human influenza a virus. *Journal of Virology*, 82(10):4807–4811, 2008.

[18] J. M. Braverman, R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, 140(2):783–96, 1995.

[19] B. Brown, A. Woerner, and J. Wilder. Ascertainment bias and the pattern of nucleotide diversity at the human ALDH2 locus in a japanese population. *Journal of Molecular Evolution*, 64(3): 375–385, 2007.

[20] E. Brunet, B. Derrida, A. H. Mueller, and S. Munier. Effect of selection on ancestry: an exactly soluble case and its phenomenological generalization. *Physical review E*, 76(4 Pt 1):041104, 2007.

[21] M. G. Bulmer. *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford, 1980.

[22] C. Cenik and J. Wakeley. Pacific salmon and the coalescent effective population size. *PLoS ONE*, 5(9):e13019, 2010.

[23] Censorinus. *De Die Natali*. 238.

[24] P. J. A. Cock et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[25] J. M. Coffin. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science*, 267(5197): 483–489, 1995.

[26] E. Cohen, D. A. Kessler, and H. Levine. Recombination dramatically speeds up evolution of finite populations. *Phys Rev Lett*, 94(9):098102, 2005.

[27] D. H. Colless. Congruence between morphometric and allozyme data for Menidia species: A reappraisal. *Systematic Zoology*, 29(3):pp. 288–299, 1980.

[28] J. M. Comeron, R. Ratnappan, and S. Bailin. The many landscapes of recombination in drosophila melanogaster. *PLoS Genet*, 8(10):e1002905, 2012.

[29] R. B. Corbett-Detig, D. L. Hartl, and T. B. Sackton. Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol*, 13(4):e1002112, 2015.

[30] C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, 1859.

[31] C. Darwin. *The Variation of Animals and Plants under Domestication*. John Murray, 1868.

[32] M. M. Desai and D. S. Fisher. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics*, 176(3):1759–98, 2007.

[33] M. M. Desai, A. M. Walczak, and D. S. Fisher. Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics*, 193(2):565–585, 2013.

[34] A. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, 7(1):214, 2007.

[35] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–7, 2004.

[36] E. E. Eichler et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*, 11(6): 446–50, 2010.

[37] B. Eldon and J. Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172(4):2621–33, 2006.

[38] W. J. Ewens. Beanbag genetics and after. In P. P. Majumder, editor, *Human Population Genetics: A Centennial Tribute to J.B.S. Haldane*, pages 7–29. Springer Science+Business Media, 1993.

[39] N. R. Faria et al. The early spread and epidemic ignition of hiv-1 in human populations. *Science*, 346(6205):56–61, 2014.

[40] J. C. Fay and C. I. Wu. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–13, 2000.

[41] J. Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78(2):737–56, 1974.

[42] C. S. Fernandez et al. Rapid viral escape at an immunodominant simian-human immunodeficiency virus cytotoxic T-lymphocyte epitope exacts a dramatic fitness cost. *J Virol*, 79(9): 5721–31, 2005.

[43] E. W. Fiebig et al. Dynamics of hiv viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary hiv infection. *AIDS*, 17(13), 2003.

[44] R. A. Fisher. *The Genetical Theory of Natural Selection*. Clarendon, 1930.

[45] Y. X. Fu and W. H. Li. Statistical tests of neutrality of mutations. *Genetics*, 133(3):693–709, 1993.

[46] V. V. Ganusov and R. J. De Boer. Estimating costs and benefits of CTL escape mutations in SIV/HIV infection. *PLoS Comput Biol*, 2(3):e24, 2006.

[47] V. V. Ganusov et al. Fitness costs and diversity of the cytotoxic T lymphocyte (CTL) response determine the rate of CTL escape during acute and chronic phases of HIV infection. *J Virol*, 85 (20):10518–10528, 2011.

[48] V. V. Ganusov, R. A. Neher, and A. S. Perelson. Mathematical modeling of escape of HIV from cytotoxic T lymphocyte responses. *J Stat Mech-Theory E*, 2013(01):P01010, 2013.

[49] N. R. Garud, P. W. Messer, E. O. Buzbas, and D. A. Petrov. Recent selective sweeps in north american Drosophila melanogaster show signatures of soft sweeps. *PLoS Genet*, 11(2): e1005004, 2015.

[50] P. J. Gerrish and R. E. Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102-103(1-6):127–44, 1998.

[51] P. J. Gerrish, A. Colato, A. S. Perelson, and P. D. Sniegowski. Complete genetic linkage can subvert natural selection. *Proc Natl Acad Sci USA*, 104(15):6266–71, 2007.

[52] J. H. Gillespie. Genetic drift in an infinite population: The pseudohitchhiking model. *Genetics*, 155(2):909–19, 2000.

[53] J. H. Gillespie. Is the population size of a species relevant to its evolution? *Evolution*, 55(11):2161–9, 2001.

[54] B. H. Good, A. M. Walczak, R. A. Neher, and M. M. Desai. Genetic diversity in the interference selection limit. *PLoS Genet*, 10 (3):e1004222, 2014.

[55] N. Goonetilleke et al. The first T cell response to transmitted/-founder virus contributes to the control of acute viremia in HIV-1 infection. *J Exp Med*, 206(6):1253–72, 2009.

[56] S. Goyal et al. Rare beneficial mutations can halt muller's ratchet. *arXiv*, q-bio.PE, 2011.

[57] A. T. Haase et al. Quantitative image analysis of HIV-1 infection in lymphoid tissue. *Science*, 274(5289):985–989, 1996.

[58] M. W. Hahn, M. D. Rausher, and C. W. Cunningham. Distinguishing between selection and population expansion in an experimental lineage of bacteriophage T7. *Genetics*, 161(1):11–20, 2002.

[59] J. Hermisson and P. S. Pennings. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4):2335–52, 2005.

[60] W. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6):226–231, 1968.

[61] W. G. Hill and A. Robertson. The effect of linkage on limits to artificial selection. *Genet Res*, 8(3):269–94, 1966.

[62] R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–8, 2002.

[63] J. Hunter. Matplotlib: a 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[64] M. Karmin et al. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Research*, 2015.

[65] B. F. Keele et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci USA*, 105(21):7552–7557, 2008.

[66] M. Kimura. Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232, 1964.

[67] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968.

[68] J. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19A:27–43, 1982.

[69] J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43, 1982.

[70] A. S. Kondrashov. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *Journal of Theoretical Biology*, 175(4):583 – 594, 1995.

[71] K. Kosheleva and M. M. Desai. The dynamics of genetic draft in rapidly adapting populations. *Genetics*, 195(3):1007–1025, 2013.

[72] M. Kreitman. The neutral theory is dead. Long live the neutral theory. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 18(8):678–683; discussion 683, 1996.

[73] E. M. Leffler et al. Revisiting an old riddle: What determines genetic diversity levels within species? *PLoS Biol*, 10(9):e1001388, 2012.

[74] S. Leviyang. Computational inference methods for selective sweeps arising in acute HIV infection. *Genetics*, 194(3):737–752, 2013.

[75] D. N. Levy, G. M. Aldrovandi, O. Kutsch, and G. M. Shaw. Dynamics of HIV-1 recombination in its natural target cells. *Proc Natl Acad Sci USA*, 101(12):4204–9, 2004.

[76] R. C. Lewontin. *The genetic basis of evolutionary change*. Columbia University Press, 1974.

[77] B. Li et al. Rapid reversion of sequence polymorphisms dominates early human immunodeficiency virus type 1 evolution. *J Virol*, 81(1):193–201, 2007.

[78] H. Li. A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Molecular Biology and Evolution*, 28(1):365–375, 2011.

[79] H. Li and T. Wiehe. Coalescent tree imbalance and a simple test for selective sweeps based on microsatellite variation. *PLoS Comput Biol*, 9(5):e1003060, 2013.

[80] M. K. Liu et al. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *J Clin Invest*, 123(1):380–393, 2013. PMID: 23221345 PMCID: PMC3533301.

[81] S.-L. Liu et al. Selection for human immunodeficiency virus type 1 recombinants in a patient with rapid progression to AIDS. *J Virol*, 76(21):10674–84, 2002.

[82] M. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits*. Sinauer, 1998.

[83] T. A. Manolio et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53, 2009.

[84] L. M. Mansky and H. M. Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol*, 69(8): 5087–94, 1995.

[85] M. Markowitz et al. A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and t-cell decay in vivo. *J Virol*, 77 (8):5037–5038, 2003.

[86] A. J. McMichael, P. Borrow, G. D. Tomaras, N. Goonetilleke, and B. F. Haynes. The immune response during acute HIV-1 infection: clues for vaccine development. *Nat Rev Immunol*, 10 (1):11, 2009.

[87] A. Mir, F. Rosselló, and L. Rotger. A new balance index for phylogenetic trees. *Mathematical Biosciences*, 241(1):125 – 136, 2013.

[88] R. A. Neher. Genetic draft, selective interference, and population genetics of rapid adaptation. *Annual Review of Ecology, Evolution, and Systematics*, 44(1):null, 2013.

[89] R. A. Neher and O. Hallatschek. Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences*, 110(2):437–442, 2013.

[90] R. A. Neher and T. Leitner. Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput Biol*, 6(1): e1000660, 2010.

[91] R. A. Neher and B. I. Shraiman. Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics*, 188(4):975–996, 2011.

[92] R. A. Neher, B. I. Shraiman, and D. S. Fisher. Rate of adaptation in large sexual populations. *Genetics*, 184:467–481, 2010.

[93] R. A. Neher, T. A. Kessinger, and B. I. Shraiman. Coalescence and genetic diversity in sexual populations under selection. *PNAS*, 110(39):15836–15841, 2013.

[94] R. Nielsen et al. Genomic scans for selective sweeps using SNP data. *Genome research*, 15(11):1566–75, 2005.

[95] S. Ohno. So much "junk" DNA in our genome. *Brookhaven Symposia in Biology*, 23:366–370, 1972.

[96] T. Oliphant. Python for scientific computing. *Comput Sci Eng*, 9 (3):10–20, 2007.

[97] R. Paredes et al. In vivo fitness cost of the m184v mutation in multidrug-resistant human immunodeficiency virus type 1 in the absence of lamivudine. *J Virol*, 83(4):2038–43, 2009.

[98] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271 (5255):1582–6, 1996.

[99] A. S. Perelson, P. Essunger, and D. D. Ho. Dynamics of HIV-1 and CD4+ lymphocytes in vivo. *AIDS*, 11 Suppl A:S17–24, 1997.

[100] J. Petravic, L. Loh, S. J. Kent, and M. P. Davenport. CD4+ target cell availability determines the dynamics of immune escape and reversion in vivo. *J Virol*, 82(8):4091–4101, 2008.

[101] J. Petravic et al. Estimating the impact of vaccination on acute simian-human immunodeficiency virus/Simian immunodeficiency virus infections. *J Virol*, 82(23):11589–11598, 2008.

[102] M. N. Price, P. S. Dehal, and A. P. Arkin. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol*, 26(7):1641–50, 2009.

[103] A. Rambaut et al. The genomic and epidemiological dynamics of human influenza a virus. *Nature*, 453(7195):615–9, 2008.

[104] A. Ramírez-Soriano and R. Nielsen. Correcting estimators of θ and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process. *Genetics*, 181(2): 701–710, 2009.

[105] A. Ramírez-Soriano, S. E. Ramos-Onsins, J. Rozas, F. Calafell, and A. Navarro. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics*, 179(1):555–567, 2008.

[106] D. P. Rice, B. H. Good, and M. M. Desai. The evolutionarily stable distribution of fitness effects. *Genetics*, 200(1):321–329, 2015.

[107] I. M. Rouzine, J. Wakeley, and J. M. Coffin. The solitary wave of asexual evolution. *Proc Natl Acad Sci USA*, 100(2):587–92, 2003.

[108] C. A. Russell et al. The global circulation of seasonal influenza A (H3N2) viruses. *Science*, 320(5874):340–346, 2008.

[109] M. J. Sackin. "Good" and "bad" phenograms. *Systematic Biology*, 21(2):225–226, 1972.

[110] J. F. Salazar-Gonzalez et al. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J Exp Med*, 206(6): 1273–89, 2009.

[111] O. Sargsyan and J. Wakeley. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theoretical Population Biology*, 74(1):104–114, 2008.

[112] T. Schwander and B. J. Crespi. Twigs on the tree of life? Neutral and selective models for integrating macroevolutionary patterns with microevolutionary processes in the analysis of asexuality. *Molecular Ecology*, 18(1):28–42, 2009.

[113] J. Schweinsberg. Coalescent processes obtained from supercritical Galton-Watson processes. *Stochastic Process. Appl.*, 106(1): 107–139, 2003.

[114] J. Schweinsberg. Rigorous results for a population model with selection II: genealogy of the population. *arXiv:1507.00394 [math]*, 2015. arXiv: 1507.00394.

[115] S. Seki and T. Matano. CTL escape and viral fitness in HIV/SIV infection. *Front. Microbio.*, 2:267, 2012.

[116] J. d. Silva. The dynamics of HIV-1 adaptation in early infection. *Genetics*, 190(3):1087–1099, 2012. PMID: 22209906.

[117] R. B. Squires et al. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and Other Respiratory Viruses*, 6(6):404–416, 2012.

[118] N. Strelkowa and M. Lässig. Clonal interference in the evolution of influenza. *Genetics*, 2012.

[119] F. Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–95, 1989.

[120] E. Talevich, B. Invergo, P. Cock, and B. Chapman. Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in biopython. *BMC Bioinformatics*, 13(1):209, 2012.

[121] L. S. Tsimring, H. Levine, and D. A. Kessler. RNA virus evolution via a fitness-space model. *Phys. Rev. Lett.*, 76:4440–4443, 1996.

[122] A. M. Walczak, L. E. Nicolaisen, J. B. Plotkin, and M. M. Desai. The structure of genealogies in the presence of purifying selection: A "fitness-class coalescent". *Genetics*, 190:753–779, 2012.

[123] D. B. Weissman and N. H. Barton. Limits to the rate of adaptive substitution in sexual populations. *PLoS Genet*, 8(6):e1002740, 2012.

[124] D. B. Weissman, M. M. Desai, D. S. Fisher, and M. W. Feldman. The rate at which asexual populations cross fitness valleys. *Theor Pop Bio*, 75(4):286–300, 2009.

[125] J. O. Wertheim and M. Worobey. Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Comput Biol*, 5 (5):e1000377, 2009.

[126] J. Yang et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 42(7):565, 2010.

[127] G. U. Yule. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 213(402-410): 21–87, 1925.

[128] F. Zanini and R. A. Neher. FFPopSim: an efficient forward simulation package for the evolution of large populations. *Bioinformatics*, 28(24):3332–3333, 2012.

[129] K. Zeng. A coalescent model of background selection with recombination, demography and variation in selection coefficients. *Heredity*, 110(4):363–371, 2013.

[130] K. Zeng, Y.-X. Fu, S. Shi, and C.-I. Wu. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, 174(3):1431–9, 2006.

[131] W. Zhai, R. Nielsen, and M. Slatkin. An investigation of the statistical power of neutrality tests based on comparative and population genetic data. *Mol Biol Evol*, 26(2):273–83, 2009.

[132] S. Zhu, C. Than, and T. Wu. Clades and clans: a comparison study of two evolutionary models. *Journal of Mathematical Biology*, 71(1):99–124, 2015.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/