Distortions in Genealogies due to Purifying Selection

Distortions in Genealogies due to Purifying Selection

A dissertation presented

by

Lauren Elisabeth Nicolaisen

to

The Department of Physics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Physics

Harvard University

Cambridge, Massachusetts

April 2014

Thesis advisor
Michael M. Desai

Author
Lauren Elisabeth Nicolaisen

Distortions in Genealogies due to Purifying Selection

Abstract

As deleterious variants continually arise in a population, they tend to be purged via purifying selection, leading to distortions in the shapes of genealogies relative to neutral expectations. In recent years, a mounting body of evidence has arisen suggesting that this can have significant implications for the patterns of diversity seen in natural populations. However, existing theory has not yet fully characterized the effects of these distortions on the structure of genealogies. The focus of this thesis is on exploring this gap, and developing an analytical description of the distortions that arise in genealogies due to purifying selection.

In the first half of this thesis, we develop a framework for calculating a variety of statistics that describe sequence variation in the strong selection regime. We will derive these results using two complementary frameworks: First, using a Poisson Random Field model to describe lineage frequencies within fitness classes, and second, using a direct extension of the structured coalescent model. In addition to enabling us to develop an analytical understanding of a number of important statistics, this will provide an intuitive picture of the nature of the distortions that arise. In particular, we show how the concept of a time-dependent effective population size emerges naturally from the structured coalescent framework.

In the latter half of this thesis, we return to our discussion of a time-dependent effective population size. We develop a method for explicitly calculating the form of this function, $N_e(t)$, as well as the analogous time-dependent effective mutation rate, $U_e(t)$. In addition, we show how this result can be extended to incorporate a variety of additional scenarios, such as recombination and a distribution of fitness effects. Within the strong purifying selection regime, this result allows us to completely describe the shapes of genealogies using a neutral framework with the appropriate $N_e(t)$ and $U_e(t)$, completely bypassing the need to model the effects of selection directly.

This implies that all of the findings of the standard neutral coalescent will still apply, and provides a simple way to incorporate purifying selection into neutral methods of inference and estimation.

Contents

List of Figures

List of Figures

Citations to Previously Published Work

Chapter 2 appears in its entirety in:

> "The Structure of Genealogies in the Presence of Purifying Selection: A Fitness-Class Coalescent", A. M. Walczak*, L.E. Nicolaisen*, J.B. Plotkin, and M.M. Desai, Genetics **190**, 2 (2012).

Chapter 3 appears in its entirety in:

> "The Structure of Allelic Diversity in the Presence of Purifying Selection", M. M. Desai*, L. E. Nicolaisen*, A. M. Walczak, and J. B. Plotkin, Theoretical Population Biology **81**, 2 (2012).

Chapter 4 appears in its entirety in:

> "Distortions in Genealogies due to Purifying Selection", L. E. Nicolaisen and M. M. Desai, Molecular Biology and Evolution **29**, 11 (2012).

Chapter 5 appears in its entirety in:

> "Distortions in Genealogies due to Purifying Selection and Recombination", L. E. Nicolaisen and M. M. Desai, Genetics **195**, 1 (2013).

* denotes equal authorship

Acknowledgments

It's a unique and adventurous thing to join a lab that is just starting up, as it's nearly impossible to predict exactly how things will turn out. I think I can safely say at this point, however, that joining Michael Desai's lab was by far the best decision I made in graduate school. Michael had a near-perfect blend of extensively helping and mentoring me early on, yet allowing me essentially complete freedom by the end. Throughout, it was always clear that he would support me and allow me to pursue any problem or question I found interesting. I couldn't have asked for a better advisor.

I am also fortunate enough to have an unbelievably great committee. I distinctly remember when I gave my first presentation outside of the safe confines of my lab group – David Nelson was in the audience and kept asking questions with such interest and enthusiasm that I felt completely at ease. By the end of that talk, his enthusiasm was so contagious that I was convinced that my research topic was the most interesting of all possible topics in the universe.

John Wakeley literally taught me everything I know about coalescent theory. In fact, much of the inspiration for this thesis came out of a series of revelations following my final project for his class. His support in reading early drafts of each of the papers in this thesis provided a huge amount of motivation and self-confidence, and I was truly lucky to have him on my committee.

I owe a huge debt of gratitude to everyone in my lab, and throughout the Physics department, as well as to Andrew Murray and all of the members of the Murray lab. There was never a time when I didn't have someone to talk to, someone to bounce ideas off of, someone to read my early drafts. I owe a special thanks to Ben Good – I couldn't possibly describe how amazing it is to have a friend around who knows your research well and is willing to sit and talk with you about it for hours on end. This thesis wouldn't be nearly as complete without his immense help along the way.

Throughout graduate school, I had an amazing group of close friends. Two in particular deserve my endless, impossible-to-quantify thanks. Sebastian, my first permanent office-mate, who knew all of my many flaws and stuck by me nonetheless, who supported me when I most needed it. And to Max, who deserves a better thanks than I'll be able to come up with for being a truly great friend.

My path into physics was anything but a straight one – there were literally dozens

of people along the way who just happened to say the right thing at the right time, nudged me in the right direction, supported my decisions, and just generally guided me to this path. A few people in particular made a huge impression on me: To Mrs. Griffin, for first teaching me that I loved math in 7th grade, and to Mr. Dick, for reminding me of that years later when I'd all but forgotten. To Francoise Queval for trusting me when I said I wanted to be a physicist, and to Prof. Mark Morris for giving me a chance to get started. I couldn't possibly thank you all enough.

And lastly, I have just about the best family in the world. There's nothing I could say that would do justice to everything they've done for me. To my parents, for taking me on endless adventures as a kid and showing me the world, for supporting and loving me every single minute of my life. And to my two incredible brothers: Chris, for constantly reminding me that I'd never be alone no matter how far away I moved, and to Dan, for being the best big brother anyone could dream of asking for. You all mean the world to me.

To Mom, Dad,
Dan, and Chris.

Chapter 1

Introduction

Over thirty years ago now, the first analysis of genetic variation at the nucleotide level was performed (Kreitman 1983). Not long after came the first data-driven analyses of genealogies and gene trees (for example, Stephens and Nei (1985), Aquadro *et al.* (1986)). With the advent of PCR and the improvement of sequencing technologies, studies with more and more individuals, covering longer and longer sequence lengths came about. The cost of DNA sequencing decreased faster than exponentially, and today, with the development of next-generation sequencing technologies and concerted efforts to sequence large samples from populations across their genomes, there is simply a staggering amount of DNA sequence data available, and it continues to grow at a rapid rate (see Powell (1994) and Charlesworth (2010) for reviews of the early history of molecular genetic techniques).

With this data comes the fundamental question, which is central to population genetics: What exactly can we discern about the history of a population from this data? As our ability to generate massive amounts of experimental data has grown tremendously over the years, so too has our understanding of population genetics from a theoretical perspective. Today, there exists a deep and insightful literature on the study of gene genealogies and the retrospective analysis of samples from a population. This understanding has provided a wealth of bioinformatic methods and tools that allow us to analyze the patterns of DNA sequence data, and to reach conclusions about the history of populations. However, our abilities are still limited in many

ways. A significant number of open questions remain, and the effects and patterns expected under a wide variety of complicated scenarios are still poorly understood. Furthermore, even in cases where we have a solid conceptual understanding of the patterns we expect to see, there are often no practical methods for detecting or analyzing these patterns in practice.

The focus of this thesis is on one particular area of theoretical population genetics: the effects of purifying selection, that is, the continuous creation and removal of deleterious variants from a population. We will explore the effects of purifying selection on genealogies, and analyze the patterns and distortions we expect to see as a result. Finally, we will investigate methods for predicting and detecting these patterns in practice. However, before we can delve into the inner workings of purifying selection and genealogies, we have to start from the beginning, by introducing one of the first, foundational models in population genetics, the Wright-Fisher model.

1.1 The Wright-Fisher Model

The foundations of modern theoretical population genetics were laid out in the 1920s and early 1930s in pioneering works by Wright, Fisher, and Haldane (Wright 1931; Fisher 1930; Haldane 1927). During this time, the widely-used stochastic framework known as the Wright-Fisher model was developed. In the simplest version of this model, we can imagine a population of constant size N. Each generation, all of the individuals in the population will die and be replaced by their offspring. The offspring are chosen from the previous generation via random sampling with replacement. In other words, each descendant will have an ancestor (parent) randomly chosen from the previous generation. This is shown schematically in Figure 1.1.

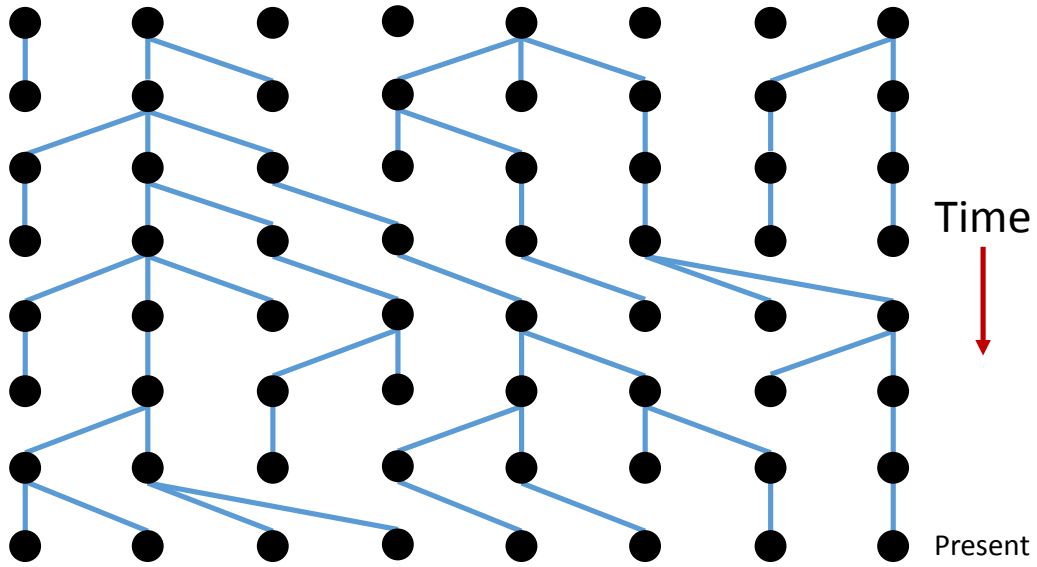

Figure 1.1: **Schematic of the Wright-Fisher Model:** Each generation is chosen from the previous generation via random sampling with replacement.

We are interested in examining the consequences for this model on the expected diversity in the population. Suppose, for example, that individuals in this population may have one of two allelic types at a single locus, denoted either A or a. In the current generation, we label the total number of individuals of type A as i, such that the fraction of individuals is $p = i/N$. We are interested in the distribution of the number of individuals of type A in the next generation, which is given by the binomial formula:

$$P(j) = \binom{N}{j} p^j (1-p)^{N-j}. \tag{1.1}$$

From this, we immediately see that:

$$\begin{aligned} E[j] &= Np \\ \mathrm{Var}[j] &= Np(1-p). \end{aligned}$$

We note that this equation depends only upon the current fraction of the population in a given allelic state, and thus is independent of all previous generations. Therefore, we can describe the forward-time dynamics of the population as a Markov process

with transition probabilities given by $P(j|i)$. From this starting point, a vast number of major results in theoretical population genetics can be derived (Ewens 2004).

The Wright-Fisher process is an inherently stochastic process, such that the frequencies of alleles in the population are subject to extensive random fluctuations. These random fluctuations are collectively referred to as genetic drift, which operates on a time-scale of $1/N$. To get a sense for this time-scale, we can consider a simple concrete example (from Wakeley (2009)). Suppose we sample two individuals from a population at random: What is the probability that they are of the same allelic type, i.e., what is the heterozygosity? In the present generation, this is simply:

$$H = 2p(1-p). \tag{1.2}$$

However, in order to calculate the average heterozygosity in the next generation, we have that:

$$\begin{aligned} H' &= E[2p'(1-p')] = 2E[p'] - 2E[p']^2 - 2\mathrm{Var}[p'] \\ H' &= 2p - 2p^2 - 2p(1-p)/N = 2p(1-p)(1-1/N) \\ H' &= H\left(1 - \frac{1}{N}\right). \end{aligned}$$

Therefore, in subsequent generations, we have that the average heterozygosity is:

$$H(t) = H(0)\left(1 - \frac{1}{N}\right)^t \approx H(0)e^{-t/N}. \tag{1.3}$$

Thus, we see that the average heterozygosity is decreasing with time, on a time-scale of order $1/N$. Eventually, all individuals in the population will be of the same allelic class, and the heterozygosity in the population will fall to zero. This will occur when one of the two alleles fixes in the population, and thus the other allele will have died out (i.e. when one of the two absorbing states of the Markov process are reached).

Related to this, we may also be interested in the probability that the allele to fix is that of type A, given that there are initially i individuals of type A. Using the transition probabilities from above, and denoting $f(i)$ as the probability of fixation starting with i individuals, we have that:

$$f(i) = \sum_{j=0}^{N} f(j)P(j|i), \tag{1.4}$$

Recall that $E[j|i] = \sum_{j=0}^{N} jP(j|i) = i$, such that we can immediately see that the solution to the above system of equations is simply a constant times the initial frequency. Using the boundary condition $f(N) = 1$, we see that $f(i) = \frac{i}{N}$.

An even simpler way to derive this result is to note that, eventually, all individuals in the population will be descended from one particular ancestor in the present. Since all individuals in the present generation are equivalent in their distribution of offspring number, the probability that any particular individual is the ancestor is simply $1/N$, and thus the probability that the ancestor is one of the i individuals is i/N. This concept of equivalent reproductive potential is intimately connected with another key concept: that of exchangeability.

1.1.1 Approximations and Exchangeability

We have seen that the Wright-Fisher model is a very powerful tool for describing the dynamics of a population. However, the model makes a number of key simplifications. First, we assume that the size of the population is constant over time and that there is no recombination. We also assume that there is no geographic structure, nor is there any selection. These last two points are part of a more general statement about the Wright-Fisher process: all lineages within the population are entirely exchangeable. This implies that the distribution of offspring number for all individuals in the population is identical, and there can be no 'labeling' of individuals, nor can there be any transmission of labels over generations (Wakeley 2009).

This last point will be of key importance in this thesis. When purifying selection operates, individuals that contain deleterious mutations are less-fit, and thus less likely to produce offspring. This violates the assumption of exchangeability, and prevents us from using these simple results to describe the effects of purifying selection directly. Instead, we will need to rely on an expanded model that allows us to incorporate this non-exchangeability into our framework.

1.1.2 A Retrospective vs. Prospective Approach

Thus far, we have been analyzing the dynamics of populations *prospectively*, that is, forwards-in-time. This requires us to describe the complete dynamics of the entire population throughout its history. However, in practice, we are often interested only in the history of a subset of individuals from a population, sampled in the present. One of the key insights of theoretical population genetics in the 20th century was to recognize that we can instead analyze samples *retrospectively*, that is, backwards-in-time, using a novel approach known as the coalescent.

1.2 The Coalescent

A common scenario in experimental population genetics is for a sample of individuals from a larger population to be sequenced, with the goal of using the observed patterns of molecular diversity to make statistical inferences about the history of that population. Although the prospective, forward-time, approach from the previous section allows us to draw a number of conclusions about this diversity, it is often far simpler to understand these patterns by taking a retrospective, backwards-in-time approach.

To do this, we will start by considering the genealogy, or gene tree, of a sample. To describe a genealogy, we trace the ancestral lineages of our sample backwards-in-time. At some point in the past, two of the ancestors of our sample will descend from the same parent. At this point, the two lineages fuse together into one lineage, which is termed coalescence.

The genealogy is then depicted as a bifurcating tree, with time running vertically from the top to the bottom (present). At the base of the tree, there are n distinct individuals, representing the sample taken in the present. At each coalescent event, the number of distinct lineages decreases by 1, to $n-1$, then $n-2$, etc. until there is only one remaining lineage in the ancestry. As an example, consider again the population depicted in Fig. 1.1. If we sample five individuals from this population, which we have marked red, we can trace the ancestral relationships between these individuals, also in red, to reconstruct the tree shown in Fig. 1.2.

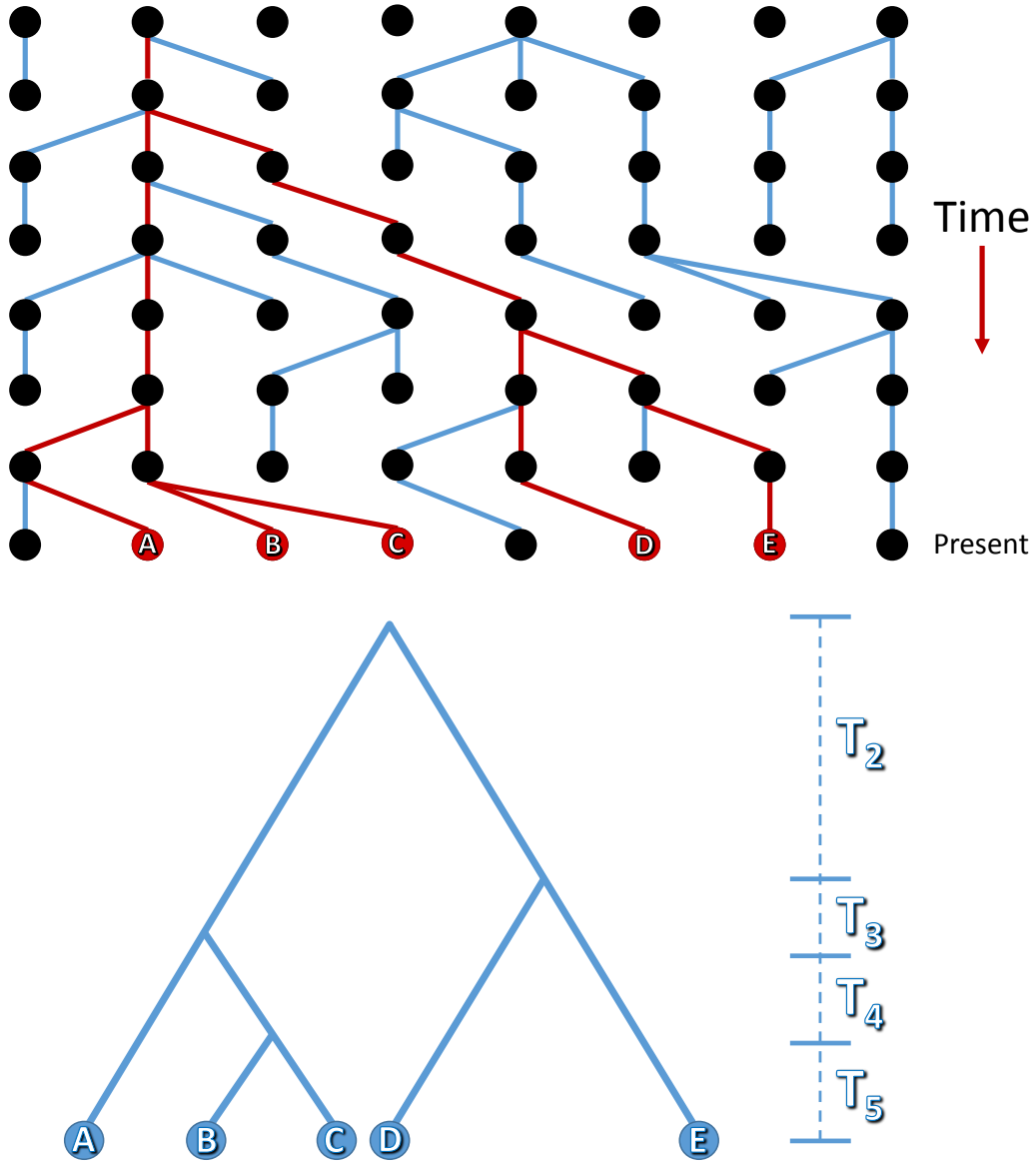

Figure 1.2: **Schematic of a Genealogy:** A sample of five individuals is chosen from the population, highlighted in red. Their ancestral histories are then traced backwards-in-time, also highlighted in red. The resulting genealogy is then reconstructed in Figure 1.2b.

Chapter 1

The idea of considering the ancestry of a sample retrospectively appeared in a number of early works pre-dating the first formal analysis of coalescent theory. A few examples of this include early analyses on the concept of identity by descent (Malécot 1941), the development of early estimators such as Watterson's estimator (Watterson 1975), and a retrospective derivation of the Ewens sampling formula (Karlin and McGregor 1972). However, the first formal descriptions of the coalescent were developed in the early 1980s by Kingman (1982), as well as Hudson (1983) and Tajima (1983) (see Hudson *et al.* (1990) for an excellent review of early coalescent theory, and Nordborg (2001) and Wakeley (2009) for more recent overviews).

Coalescent theory provides a mathematical framework for completely describing the probabilities of particular gene trees. A gene tree consists of a set of ancestral relationships between each lineage, as well as a set of times at which coalescent events occurred. These times are denoted T_i, where $i = 2, 3, 4, ..., n$ represents the number of distinct lineages present in that time interval. The complete distribution of these times can be derived from a variety of forward-time models, including the Wright-Fisher model.

Consider a sample of i lineages taken from the present generation. In order for two individuals to coalesce in the previous generation, they must share the same parent. Thus, the probability that there are no coalescent events among any of the i lineages is simply the probability that all i descendants have distinct parents:

$$P(\text{no coal.}|i) = \left(\frac{N-1}{N}\right)\left(\frac{N-2}{N}\right)\ldots\left(\frac{N-i+1}{N}\right) = 1 - \frac{\binom{i}{2}}{N} + \mathcal{O}\left(\frac{1}{N^2}\right).$$

Similarly, the probability that exactly one coalescent event occurs is:

$$P(\text{one coal.}|i) = \left(\frac{\binom{i}{2}}{N}\right)\left(\frac{N-1}{N}\right)\ldots\left(\frac{N-i+2}{N}\right) = \frac{\binom{i}{2}}{N} + \mathcal{O}\left(\frac{1}{N^2}\right).$$

In the limit $N \to \infty$, the probability that two or more pairs of lineages coalesce in the same generation can be neglected, and the probability of coalescence is simply $P(\text{coal}|i) = \frac{\binom{i}{2}}{N}$. From here, the distribution of times until the first coalescent event

is given by the geometric distribution:

$$P(t_i = t) = \left(1 - \frac{\binom{i}{2}}{N}\right)^t \frac{\binom{i}{2}}{N}.$$

Typically, we will re-scale time in units of N generations, such that $T_i = t_i/N$. Thus, in the limit $N \to \infty$, this becomes:

$$P(T_i = T) \approx \binom{i}{2} e^{-\binom{i}{2}T}. \tag{1.5}$$

Therefore, we see that the time to coalescence is exponentially-distributed with rate $\binom{i}{2}$. Furthermore, we know that in the Wright-Fisher model, all individuals are completely exchangeable. As a consequence of this, each pair of lineages is equally likely to coalesce at each step and the time to coalescence at each step is independent of the time to coalescence at every other step. Together, these points allow us to calculate the complete probability of any particular gene tree.

However, an important fundamental point is that the gene trees themselves are inherently unobservable. Rather, in practice, we will observe polymorphism data, that is, the set of mutations that occur along the genealogy. However, the patterns that we see in this data are direct consequences of the shapes of the genealogies, such that the polymorphism data itself will provide us insight into the underlying genealogy, and therefore, the evolutionary process behind the genealogy.

1.2.1 Incorporating Mutations

Under the Wright-Fisher model, all individuals are unlabeled and entirely exchangeable. Thus, when a neutral mutation occurs, it has no effect on the underlying coalescent process or the shapes of genealogies. As a consequence, the mutation process can be completely separated from the coalescence process. We will typically assume that neutral mutations occur at a constant, per-generation rate of U, and that the distribution of the number of mutations that occur along a branch of length t is Poisson-distributed with mean Ut. Since time is typically scaled in units of N generations, we will typically use the scaled mutation rate, $\Theta = 2NU$, such that the number of mutations along a branch of length T is Poisson-distributed with mean

$\Theta T/2$. For example, suppose that we are interested in the distribution of the total number of mutations that occur in a sample of size $n = 2$. We have that:

$$\begin{aligned} P(S_2 = k) &= \int_0^\infty P(k|T)P(T_2 = T)dT \\ P(S_2 = k) &= \int_0^\infty \frac{(\Theta T)^k}{k!} e^{-\Theta T} e^{-T} dT \\ P(S_2 = k) &= \left(\frac{\Theta}{1+\Theta}\right)^k \left(\frac{1}{1+\Theta}\right). \end{aligned}$$

An alternate, intuitive way to derive this quantity is to recognize that both the coalescence and mutation processes are approximately Poisson processes with rates $\binom{i}{2}$ and $\frac{i\Theta}{2}$, respectively. Thus, in general, for a sample of i individuals, the probability that the next event is a mutation is simply $\frac{\Theta}{i-1+\Theta}$. In order for a sample of size $n = 2$ to have k mutations, there must be k mutation events, followed by a coalescent event. This will occur with probability:

$$P(S_2 = k) = \left(\frac{\Theta}{1+\Theta}\right)^k \left(\frac{1}{1+\Theta}\right).$$

Using this framework, we can now calculate the probability of any particular genealogy, incorporating mutations. However, as noted previously, the genealogy itself is inherently unobservable. The only information we will typically have about a population is in the form of polymorphism data. Our goal is to use this polymorphism data to try and draw inferences about the history of a population. Throughout our analysis, we will typically make use of an infinite-sites model. The infinite-sites model assumes that all new mutations occur at sites that had not previously held a mutation. When this is the case, there is a one-to-one correspondence between mutations along the genealogy and corresponding polymorphisms in the sequence data (Wakeley 2009). Thus, for example, if a sample were to have evolved according to the genealogy shown in Figure 1.3a, then the resulting observed sequence data would be analogous to that of Figure 1.3b.

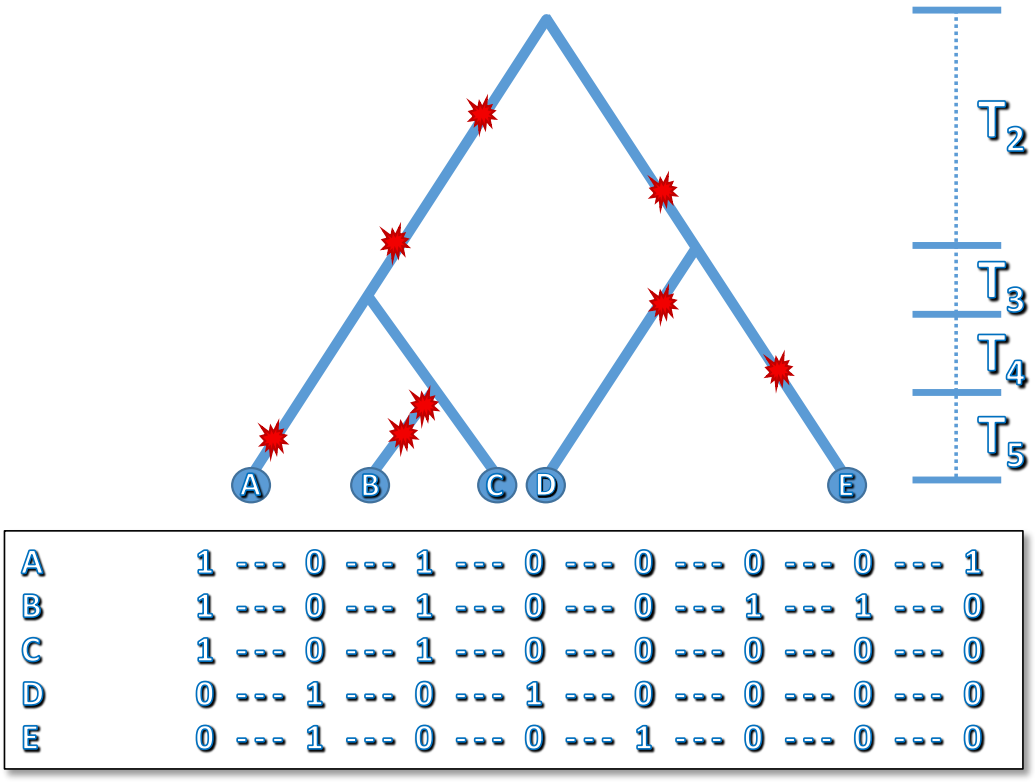

Figure 1.3: **Correspondence between Genealogies and Sequence Data:** Figure 1.3a depicts a genealogy for a sample of size 5, with mutations highlighted in red. Under the infinite-sites model, each of these mutations corresponds to a new, unique polymorphism in the population, such that the resulting sequence data will appear similar to Figure 1.3b.

1.2.2 Summary Statistics

The simplest method for inference is the use of summary statistics. The goal is to use the coalescent framework to calculate the distribution of an observable statistic. By comparing the observed value with its likelihood under our model, we can attempt to estimate the value of any unknown parameter(s). For example, we showed in the previous section that the distribution of the number of segregating sites in a sample

of size two was simply:

$$P(S_2 = k) = \left(\frac{\Theta}{1+\Theta}\right)^k \left(\frac{1}{1+\Theta}\right).$$

We can extend this analysis to arbitrary sample size. For example, the expected number of segregating sites in a sample of size n is:

$$\begin{aligned} E[S_n] &= \sum_{i=2}^{n} \frac{i\Theta}{2} E[T_i] = \sum_{i=2}^{n} \frac{i\Theta}{2} \frac{1}{\binom{i}{2}} \\ E[S_n] &= \sum_{i=1}^{n-1} \frac{\Theta}{i} = \Theta H_{n-1}, \end{aligned}$$

where H_{n-1} is the $(n-1)$th Harmonic number. One of the earliest estimators in population genetics, Watterson's estimator, makes use of this relationship to calculate a simple estimator for Θ (Watterson 1975):

$$\hat{\Theta_w} = \frac{S_n}{H_{n-1}}. \tag{1.6}$$

Another commonly used summary statistic is the average number of pairwise differences, i.e. the mean number of differences between each pair in the sample. Since all pairs of lineages are exchangeable, the expected value of this statistic is simply:

$$E[\pi] = \Theta. \tag{1.7}$$

In our example in Figure 1.3, the sample has $n = 5$, $S_5 = 8$, and $\pi = 3.8$. Although summary statistics can be very useful, the full power of the coalescent is in our ability to use this framework for more detailed inference methods. We have seen how we can use the coalescent framework to calculate the probability of a particular tree, given a set of parameters, $P(\text{tree}|\Theta)$. In practice, however, there are a large number of possible genealogies that can lead to an observed data set, and explicitly calculating $P(\text{data}|\Theta)$ requires summing over all such possible trees. In order to make this process feasible, more efficient means of sampling must be employed. Developing full-scale inference methods is a major ongoing effort in population genetics, and has led to a number of widely-used bioinformatic tools and techniques (see Tavaré (2004) and Stephens (2008) for reviews of inference in general, and Kuhner (2009)

for a review of specific technologies). Despite several major advances in recent years, these full-scale inference methods still have a number of limitations: in particular, it is exceedingly difficult to incorporate more complicated scenarios into these methods, such as the effects of purifying selection. One of the primary focuses of this thesis is on the development of a simplified description of purifying selection that allows for the effects of purifying selection to be incorporated into these pre-existing methods.

1.2.3 Tests of Neutrality

Thus far, we have focused on a simple model assuming, among other things, that there is no selection, no population structure, and a constant population size. However, in practice, these assumptions are frequently violated. In order to investigate deviations from these assumptions, we will typically use the standard neutral coalescent as a null model, and look for deviations from our predictions.

One of the earliest examples of a 'test for neutrality' is Tajima's D (Tajima 1989). In the previous section, we saw that both the expected number of segregating sites, S, and the average number of pairwise differences, π, are proportional to the scaled mutation rate, Θ. Therefore, both S and π may be used to estimate a value for Θ. If a population is evolving according to the neutral dynamics we have modeled, then over a large number of trials, we expect for the difference between these two estimates to average to zero. This forms the basis for the test statistic Tajima's D, which divides the difference in the two estimates by the standard deviation of their difference:

$$D = \frac{\pi - \frac{S_n}{H_{n-1}}}{\sqrt{\mathrm{Var}[\pi - \frac{S_n}{H_{n-1}}]}}. \tag{1.8}$$

Tajima (1989) showed how this variance could be estimated from the data and provided a set of p-values for rejecting the null hypothesis, assuming that the distribution of the statistic could be modeled as a beta distribution.

In order to understand how violations of the assumptions of the neutral model can lead to a deviation in this test statistic, it is informative to examine two hypothetical trees. Figure 1.4 depicts two different trees, each with the same sample size, $n = 5$, and the same number of segregating sites, $S_5 = 8$. However they differ significantly

in their underlying shape. In particular, the second tree has substantially elongated branch lengths in the recent past relative to those in the distant past. This leads to an excess of rare mutations, which implies a lower average number of pairwise differences. As a consequence, the relative estimates of Θ using the two summary statistics appears to be inconsistent with the neutral model, and we observe a negative value of Tajima's D.

There are several potential demographic scenarios that can lead to such a tree. Here, we will focus on two common scenarios that may arise: first, a growing population size, and second, purifying selection. First, consider a population that is experiencing rapid growth forward-in-time. When we analyze a sample backwards-in-time, the population size starts off large, such that the typical branch lengths in the recent past are very long. However, as time recedes into the past, the population size falls off, and thus the branch lengths in the distant past are shorter. Thus, this can potentially lead to a tree such as that in Figure 1.4b.

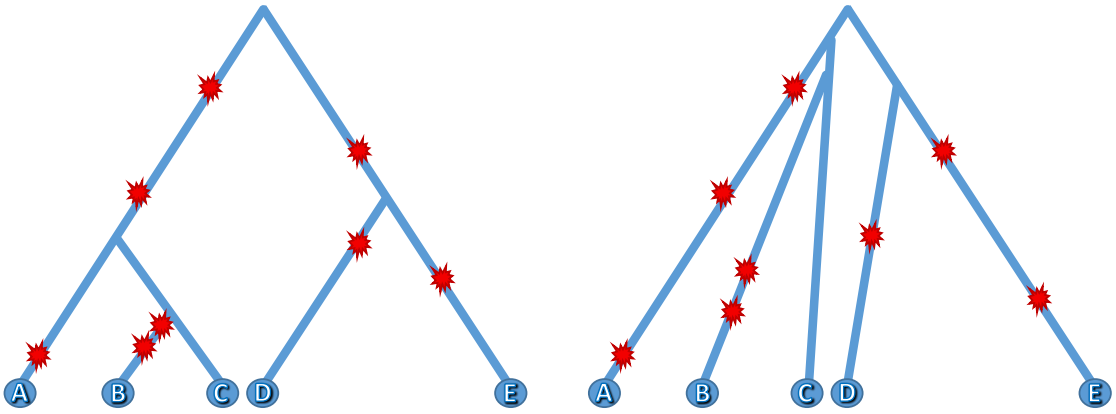

Figure 1.4: **Example of a Distorted Genealogy:** Two genealogies are shown, each with the same number of segregating sites. However, they differ significantly in their relative branch lengths. In particular, the second tree has elongated branch lengths in the recent past relative to those in the distant past, leading to a distortion in the shape of the genealogy.

However, consider instead a population that is experiencing purifying selection. Under purifying selection, individuals that acquire deleterious mutations tend to die out from the population quickly. In contrast, the more fit individuals tend to persist. Thus, as we analyze a sample backwards-in-time, ancestors become biased towards coming from the more fit individuals, and the effective population size (defined here as the inverse of the rate of coalescence) decreases with time. Note that, in addition to causing distortions in the branch lengths in the genealogy, this also causes the distribution of mutations along the genealogy to be non-neutral. In particular, since individuals with deleterious mutations tend to die out quickly, deleterious mutations will be more common in the recent past than in the distant past.

This leads to a fundamental problem – each of the causes we described above leads to very similar signals in the data. In fact, one of the primary results in this thesis, which we will later see, is that within the strong selection regime, purifying selection can be explicitly described by using a time-varying effective population size, and we will calculate the form of this function. Thus, within this regime, the two scenarios will be indistinguishable. Although we have focused here on only two potential causes, there are several other potential causes for a negative value of Tajima's D. For example, this can be caused by positive selection (e.g. a selective sweep), or even by sequencing error (which may lead to an excess of spurious singleton mutations, Pool *et al.* (2010)).

1.3 Purifying Selection

As deleterious mutations continually arise in a population, they tend to be purged by purifying selection. When selection is very strong relative to genetic drift, deleterious variants will be removed extremely rapidly, and this process is roughly instantaneous on the time-scale of coalescence. As a result, all individuals in the population are very recently descended from individuals without deleterious mutations, and thus molecular variation is equivalent to that of a neutral population with a reduced effective population size, N_e, where N_e is the average number of individuals without deleterious mutations.

This intuitive approximation (often referred to as the 'background selection' approximation, though we will use this phrase to include all effects of purifying selection on linked variation, and the more specific phrase 'fixed effective population size' approximation to refer to this particular limit) was originally developed in Charlesworth *et al.* (1993), as well as Charlesworth (1994) and Hudson and Kaplan (1994, 1995a). It has had an enormous influence on the study of purifying selection, and has been widely used to interpret patterns of molecular variation in sequence data (Hudson and Kaplan 1995b). It successfully captures the dominant effect of strong purifying selection on genealogies: an overall decrease in coalescent times.

However, even when purifying selection is strong, it does not act instantaneously. Typically, deleterious variants will segregate in the population for a time of order $1/s$. Thus, since purifying selection has not yet had time to act against recent mutations, the effective population size in the recent past is larger than in the distant past. This leads to an overall distortion in the branch lengths of the genealogy (McVean and Charlesworth (2000); Comeron and Kreitman (2002); see Charlesworth (2013) for a review). In order to understand the nature of these distortions, we will need to develop a framework to analyze the effects of purifying selection.

1.3.1 Mutation-Selection Balance

Consider a population of constant size N, where deleterious mutations can occur at a genome-wide rate of U_d, and confer some fitness advantage s. We will assume an infinite-sites model with no epistasis, such that the fitness of an individual that carries k deleterious mutations is $\omega_k = (1-s)^k \approx 1 - sk$, where we have assumed that $s \ll 1$. If we label the fraction of the population that has k deleterious mutations as h_k, then we have that:

$$h_k(t+1) = h_k(t)\frac{\omega_k(1-U_d)}{\overline{\omega}} + h_{k-1}(t)\frac{\omega_{k-1}U_d}{\overline{\omega}}.$$

Note that $\omega_0 = 1$, therefore in steady-state we know that:

$$\begin{aligned} h_0 &= h_0\left(\frac{1-U_d}{\overline{\omega}}\right) \\ &\rightarrow \overline{\omega} = 1 - U_d. \end{aligned}$$

Therefore, in steady state, keeping only first-order terms in s, we have that:

$$h_k = h_{k-1}\frac{U_d}{sk}.$$

This implies that the fraction of the population in 'fitness class' k is Poisson distributed with mean U_d/s:

$$h_k = \frac{e^{-U_d/s}}{k!}\left(\frac{U_d}{s}\right)^k. \tag{1.9}$$

This result is known as the mutation-selection balance (Kimura and Maruyama 1966; Haigh 1978). In general, this reflects the balance between two competing forces: mutation, which introduces new deleterious mutations into the population, and selection, which purges these mutations. The population will exist in a steady-state balance between these forces when the effects of selection are strong relative to the effects of genetic drift, e.g. when $Ns \gg 1$. A schematic of this distribution is shown in Fig. 1.5.

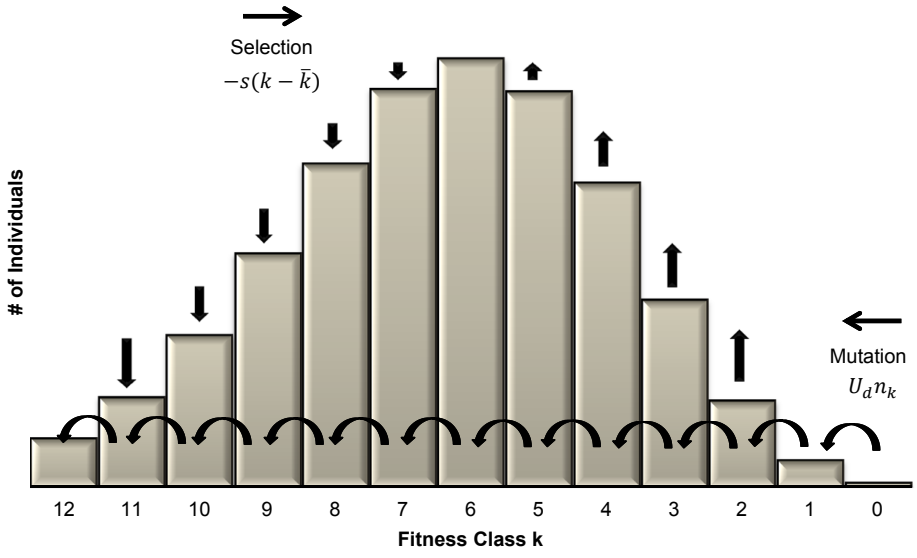

Figure 1.5: **: Schematic of Mutation-Selection Balance (from Nicolaisen and Desai (2012)):** Deleterious mutations decrease the mean fitness of the population, while selection favors more-fit individuals. At steady state, a balance between these two effects is reached.

It is informative to reconsider the 'fixed effective population size' approximation in light of this distribution. When purifying selection is very strong, all individuals are very recently descended from individuals without deleterious mutations. In other words, they are descended from individuals in the zero-class, which has size $Nh_0 = Ne^{-U_d/s}$. If we assume that the time-scale on which individuals are descended from the zero-class is effectively instantaneous on the time-scale of coalescence, then we can treat the molecular variation in the population as equivalent to that of an entirely neutral population, of size $Ne^{-U_d/s}$.

However, it is important to note that, even when selection is moderately strong and the mutation-selection balance holds, the time-scale on which individuals are descended from the zero-class is not instantaneous. Thus, in order to analyze the effects of purifying selection, we need to understand the effects that this movement through the distribution of fitness classes has on genealogies.

1.3.2 The Structured Coalescent

Consider two individuals sampled from a population, both of whom are in the same fitness class k. We will trace the ancestral lineage of these individuals backwards-in-time, as we did in the standard neutral coalescent. However, in this case, we will now also keep track of their location in the fitness distribution. There are two types of events that may occur in the ancestry of this sample. First, the two individuals may coalesce while still in fitness class k, or second, one of the two individuals may undergo a deleterious mutation (backwards-in-time) from fitness class $k-1$.

Within each fitness class, since all individuals have the same fitness, we may model the coalescent process using the neutral coalescent. Thus, we know that the rate of coalescence within fitness class k is $\frac{1}{Nh_k}$. Furthermore, we know that the rate of mutation is approximately $\frac{Nh_{k-1}U_d}{Nh_k} \approx sk$. Therefore, we have that:

$$P(\text{1st Event is Coal.}|k,k) = \frac{1/(Nh_k)}{sk+sk+1/(Nh_k)} = \frac{1}{1+2Nh_k sk} \tag{1.10}$$

$$P(\text{1st Event is Del. Mut.}|k,k) = \frac{2sk}{sk+sk+1/(Nh_k)} = \frac{2Nh_k sk}{1+2Nh_k sk}. \tag{1.11}$$

In contrast, if the two individuals are in different fitness classes, k and k', the only event that may occur is a mutation. However, it may occur in either lineage. Thus, we have that:

$$P(\text{1st Event is Del. Mut. in } k|k,k') = \frac{sk}{sk+sk'} \tag{1.12}$$

$$P(\text{1st Event is Del. Mut. in } k'|k,k') = \frac{sk'}{sk+sk'}. \tag{1.13}$$

In this manner, we can trace the ancestral lineages through the fitness class distribution. In general, this is equivalent to treating the population as though it is subdivided into the different fitness classes, but assuming that the neutral coalescent holds within each fitness class. This requires assuming that the size of each fitness class is sufficiently large to neglect fluctuations, which is roughly true provided $Nse^{-U_d/s} \gg 1$.

This framework, known as the structured coalescent, was first developed for describing selection in Hudson and Kaplan (1994, 1995a) (see Wakeley (2010) for a review of this and other frameworks for describing selection). It has formed the basis for several numerical and simulation-based studies (Zeng and Charlesworth 2011; Gordo *et al.* 2002), as well as provided a solution for the effects of selection on a single site (Barton and Etheridge 2004). A schematic of this framework is shown in Fig. 1.6. In this thesis, we will expand upon this framework to develop new methods for describing the effects of purifying selection that allow us to directly calculate the analytical effects of selection at many linked sites.

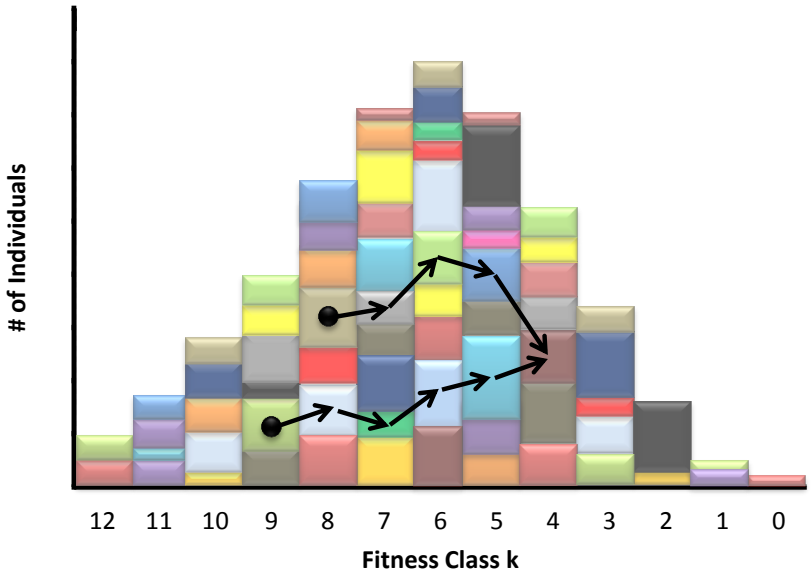

Figure 1.6: **Schematic of the Structured Coalescent (from Desai *et al.* (2012)):** Each fitness class in the population is composed of many lineages, each of which was created by a single mutation from the previous class. The arrows denote an example of the coalescence process for two individuals sampled from the population.

1.4 Outline of Thesis and Summary of Major Results

In recent years, mounting experimental evidence has arisen suggesting that selective forces are of fundamental importance for our understanding of natural populations (Hahn 2008; Comeron *et al.* 2008; Seger *et al.* 2010). However, existing theory has not yet fully characterized the effects of purifying selection on the structure of genealogies. The focus of this thesis is on filling this gap, and developing an analytical description of the distortions that arise in genealogies due to purifying selection.

We begin in Chapter 2 by developing a framework for calculating a variety of statistics that describe sequence variation, most notably, the distribution of the times to coalescence between two individuals, and the distribution of the number of neutral and deleterious mutations between those individuals. These results are found using two complementary frameworks: first, by using a Poisson Random Field description of the allele frequencies within fitness classes, and second, using a direct extension of the structured coalescent framework presented above.

Although the analytical findings presented in Chapter 2 are potentially of major practical significance, perhaps the most important aspect of the results shown therein is in the intuitive picture that arises. In particular, we find that the effect of negative selection is similar to that of an effective population size that declines as time recedes into the past. Although this analogy has been presented in earlier work (Williamson and Orive 2002; Seger *et al.* 2010), we show how this phenomenon can be extracted naturally from the framework. However, a key point of Chapter 2 is that this is not the only effect that arises from the framework: in addition to the distortions that arise due to a time-dependent effective population size, there are also topological distortions in the genealogies. This stems from the fact that lineages are no longer exchangeable: the probability of coalescence at later times depends upon the history of coalescence at earlier times. Thus, the time-dependent effective population size that is derived for a sample of size 2 will not necessarily extend to larger sample sizes, and thus does not provide a complete description of the shape of genealogies.

In Chapter 3, we continue our analysis in the vein of Chapter 2 to describe the structure of allelic diversity for a population undergoing purifying selection. In par-

ticular, we derive an analog to the Ewens sampling formula (Ewens 1972) in the case of purifying selection. As with our findings from Chapter 2, we again conclude that there is a distortion due to purifying selection, and we are able to analytically describe the shape of these distortions.

In Chapter 4, we return to our discussion of a time-dependent effective population size, $N_e(t)$. We saw previously that purifying selection has two main effects on the shapes of genealogies: first, it leads to a distortion in the relative branch lengths within the genealogy, and second, it leads to topological distortions due to the fact that lineages are no longer exchangeable. However, in the strong purifying selection regime, when $N_e s \gg 1$, the latter effects may be sufficiently small that they can be neglected. In this regime, we can describe the effects of purifying selection using a time-dependent effective population size, $N_e(t)$, which we are able to calculate explicitly in Chapter 4. Furthermore, we are able to calculate an analogous time-dependent effective mutation rate, $U_e(t)$.

There are significant implications to this finding: this allows us to completely describe the shapes of genealogies using only an $N_e(t)$ and $U_e(t)$, completely bypassing the need to model the effects of selection directly. This implies that all of the findings of the standard neutral coalescent, including the assumption of exchangeability, still holds, and thus enables us to calculate any statistic of interest using the neutral framework. Furthermore, this provides a simple way to incorporate purifying selection into neutral methods of inference and estimation.

In Chapter 5, we show how this result can be extended to incorporate additional scenarios, including the effects of recombination, a (real) time-varying population size, and a distribution of fitness effects. Thus, our findings allow us to understand the effects of strong purifying selection in a variety of situations, and provide a simple and intuitive way to incorporate selection into neutral methods of inference and estimation.

Chapter 2

The Structure of Genealogies in the Presence of Purifying Selection: A Fitness-Class Coalescent

Compared to a neutral model, purifying selection distorts the structure of genealogies and hence alters the patterns of sampled genetic variation. Although these distortions may be common in nature, our understanding of how we expect purifying selection to affect patterns of molecular variation remains incomplete. Genealogical approaches such as coalescent theory have proven difficult to generalize to situations involving selection at many linked sites, unless selection pressures are extremely strong. Here, we introduce an effective coalescent theory (a "fitness-class coalescent") to describe the structure of genealogies in the presence of purifying selection at many linked sites. We use this effective theory to calculate several simple statistics describing the expected patterns of variation in sequence data, both at the sites under selection and at linked neutral sites. Our analysis combines a description of the allele frequency spectrum in the presence of purifying selection with the structured coalescent approach of Kaplan *et al.* (1988), to trace the ancestry of individuals through the distribution of fitnesses within the population. We also derive our results

using a more direct extension of the structured coalescent approach of Hudson and Kaplan (1994). We find that purifying selection leads to patterns of genetic variation that are related but not identical to a neutrally evolving population in which population size has varied in a specific way in the past.

2.1 Introduction

Purifying selection acting simultaneously at many linked sites ("background selection") can substantially alter the patterns of molecular variation at these sites, and at linked neutral sites (Hill and Robertson 1966; Kaplan *et al.* 1988; Hudson and Kaplan 1994, 1995b; McVean and Charlesworth 2000; Gordo *et al.* 2002; Seger *et al.* 2010; O'Fallon *et al.* 2010). In recent years, evidence from sequence data points to the general importance of these selective forces among many linked variants in microbial and viral populations, and on short distance scales in the genomes of sexual organisms (Hahn 2008; Comeron *et al.* 2008; Seger *et al.* 2010). In these situations, existing theory does not fully explain patterns of molecular evolution (Hahn 2008).

It is difficult to incorporate negative selection at many linked sites into genealogical frameworks such as coalescent theory, because these frameworks typically rely on characterizing the space of possible genealogical trees *before* considering the possibility of mutations at various locations on these trees. When selection operates, the probabilities of particular trees cannot be defined independently of the mutations, and the approach breaks down (Wakeley 2009; Tavaré 2004).

Despite this difficulty, a number of productive approaches have been developed to predict how negative selection influences patterns of molecular variation and to infer selection pressures from data. Charlesworth *et al.* (1993) introduced the background selection model and showed that strong purifying selection reduces the effective population size relevant for linked neutral sites (Charlesworth 1994; Charlesworth *et al.* 1995). However, weaker selection also distorts patterns of variation, in a way that cannot be completely described by a neutral model with any

effective population size (McVean and Charlesworth 2000; Comeron and Kreitman 2002), a phenomenon often referred to as Hill-Robertson interference (Hill and Robertson 1966). Several theoretical frameworks have been developed to analyze this situation. The ancestral selection graph of Neuhauser and Krone (1997) and Krone and Neuhauser (1997) provides an elegant formal solution to the problem, but unfortunately it requires extensive numerical calculations (Przeworski *et al.* 1999). These limit the intuition we can draw from this method, and make it impractical as the basis for inference from most modern sequence data. An alternative approach is based on the structured coalescent, and views the population as subdivided into different fitness classes, tracing the genealogies of individuals as they move between classes. This approach was first introduced by Kaplan *et al.* (1988) and further developed by Hudson and Kaplan (1994, 1995b) in the case where fluctuations in the size of each fitness class can be neglected. This structured coalescent approach has been been the basis for computational methods developed by Gordo *et al.* (2002), Seger *et al.* (2010), and Zeng and Charlesworth (2011), and analytical approaches such as those of Barton and Etheridge (2004), Hermisson *et al.* (2002) and O'Fallon *et al.* (2010).

In this paper, we build on the structured coalescent framework by introducing the idea of a "fitness-class coalescent." Rather than considering the coalescence process in real time, we treat each fitness class as a "generation" and trace how individuals have descended by mutations through fitness classes, moving from one "generation" to the next by subsequent mutations. We show that the coalescent probabilities in this fitness-class coalescent can be computed using an approach based on the Poisson Random Field method of Sawyer and Hartl (1992), or equivalently can be derived as an extension of the structured coalescent approach of Hudson and Kaplan (1994).

Our fitness-class coalescent theory can be precisely mapped to a coalescence theory in which certain quantities (e.g. coalescence times) have different meanings than in the traditional theory. We can then invert this mapping to determine the structure of genealogies and calculate statistics describing expected patterns of genetic variation. This approach requires certain approximations, but it also has several advantages. Most importantly, we are able to derive relatively simple analytic expressions for

coalescent probabilities and distributions of simple statistics such as heterozygosity. Consistent with earlier work, we find that the effects of purifying selection are broadly similar to an effective population size that changes as time recedes into the past. Our analysis makes this intuition precise and quantitative: we can compute the exact form of this time-varying effective population size, as defined by the rate of pairwise coalescence. We also show that this intuition has important limitations: for example, different pairs of individuals have different time-varying effective population size histories, meaning that in principle it is possible to distinguish selection from changing population size. Our approach also makes it possible to calculate the diversity of selected alleles themselves, which may be important when selection is common (Williamson and Orive 2002).

We begin in the next section by describing the fitness-class coalescent idea which underlies our approach. We then describe the details of our model and analyze two ways to implement the fitness-class coalescent. The first relies on the Poisson Random Field method of Sawyer and Hartl (1992) to describe the frequency distribution of distinct lineages within each fitness class. We show how this lineage structure can be used to compute coalescence probabilities in each fitness class. The second approach is based on tracing the ancestry of individuals in the order that events occur as described by Hudson and Kaplan (1994), and implemented numerically by Gordo *et al.* (2002). We show how we can sum over all possible ancestral paths to compute equivalent coalescence probabilities in each fitness class. The two approaches provide different and complementary intuitive pictures of the process, and depend on various approximations in somewhat different ways.

After computing coalescence probabilities with both approaches, we show how these probabilities can be used to analyze the structures of genealogies, and we calculate various statistics describing genetic variation in these populations, which we compare to numerical simulations. We then discuss the relationship between our results, neutral theory, and earlier work on selection, and we explore how various approximations limit our approach. The most important of these approximations is that we neglect fluctuations in the size of each fitness class, analogous to earlier work (Hudson and Kaplan 1994), which restricts our analysis to the case of strong se-

lection (relative to inverse population size). This approximation also means that we neglect Muller's ratchet. We describe this and related approximations and describe their regime of validity in the Discussion. Finally, in the Appendices we explore these approximations in more detail and describe how they inform the relationship between our work and earlier approaches.

2.2 The Fitness-Class Coalescent

In this section, we outline the main ideas underlying our fitness-class coalescent approach. We begin our analysis by considering the balance between mutations at many linked sites and negative selection against the mutants, which leads to an equilibrium distribution of fitnesses within a population (Haigh 1978). We illustrate this in Fig. 2.1, for the case in which all deleterious mutations have the same fitness cost. Each individual is characterized by the number k of deleterious mutations it contains. Each fitness class k contains many distinct lineages, each of which arose from deleterious mutations in more-fit individuals, as illustrated in Fig. 2.2. Neutral mutations also occur, but we consider these later.

Hudson and Kaplan (1994) observed that individuals move between fitnesses by deleterious mutations, and that when two individuals are in the same fitness class they could be from the same lineage and hence coalesce. Our fitness-class coalescent exploits this observation to define an effective genealogical process that completely bypasses the ancestral process in real time. Instead, we treat each fitness class as a "generation," and we count time in deleterious mutations: each deleterious mutation moves us from one "generation" to the next. In this way, we can trace the ancestry of individuals through the fitness distribution. For example, there is some probability that two individuals chosen from fitness class k are genetically identical (i.e. come from the same lineage). If not, they each arose from mutations within fitness class $k-1$. If both those mutations occurred in individuals in the same lineage in fitness class $k-1$, we say the two individuals "coalesced" in class $k-1$. If not, they came from different mutations from class $k-2$, and could have coalesced there, and so on. In this way, we can construct a fitness-class coalescent tree describing the relatedness

of two individuals, as illustrated in Fig. 2.2.

In this paper we show that the probability that two randomly chosen individuals who are currently in fitness classes k and k' coalesce in class $k-\ell$, $P_c^{k,k'\to k-\ell}$, is approximately

$$P_c^{k,k'\to k-\ell} = \frac{1}{2n_{k-\ell}s_{k-\ell}} A_\ell^{k,k'}, \tag{2.1}$$

where n_k is the population size of fitness class k, s_k is an effective selection pressure against these individuals, and

$$A_\ell^{k,k'} = \frac{\binom{k'}{k-\ell}\binom{k}{k-\ell}}{\binom{k+k'}{2\ell+k'-k}}. \tag{2.2}$$

This coalescent probability is inversely proportional to the population size of the fitness class, $n_{k-\ell}$, and the effective selection coefficient within that class, $s_{k-\ell}$, modified by the combinatoric coefficient $A_\ell^{k,k'}$. As we will see, this has a clear intuitive

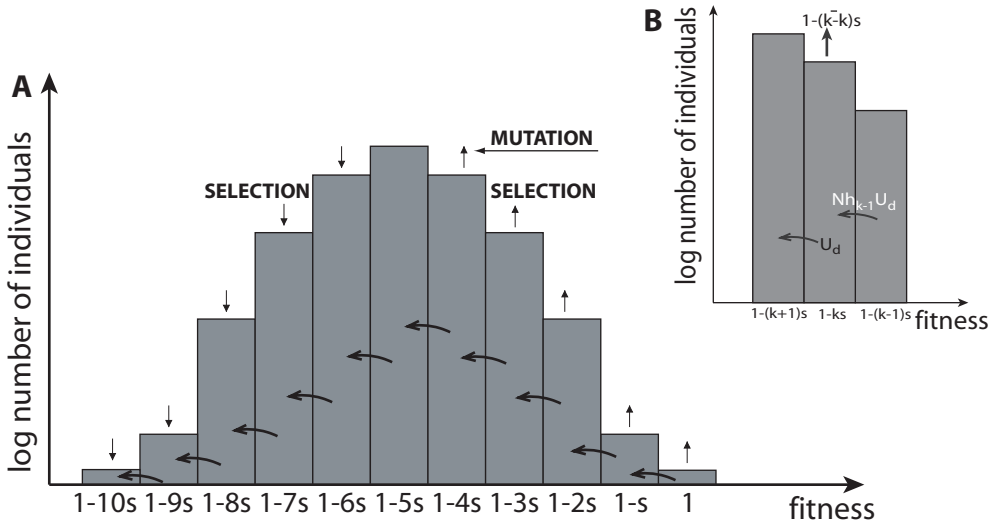

Figure 2.1: **The Distribution of the Fraction of the Population in each Fitness Class: (a)** The distribution of the number of individuals as a function of fitness, where the most beneficial class is arbitrarily defined to have fitness 1, and each deleterious mutation introduces a fitness disadvantage of s. Mutations move individuals to less-fit classes, and selection balances this by favoring the classes more fit than average. The shape of the depicted steady state distribution is a result of this mutation–selection balance. The inset **(b)** shows the processes which lead to this balance within a given fitness class.

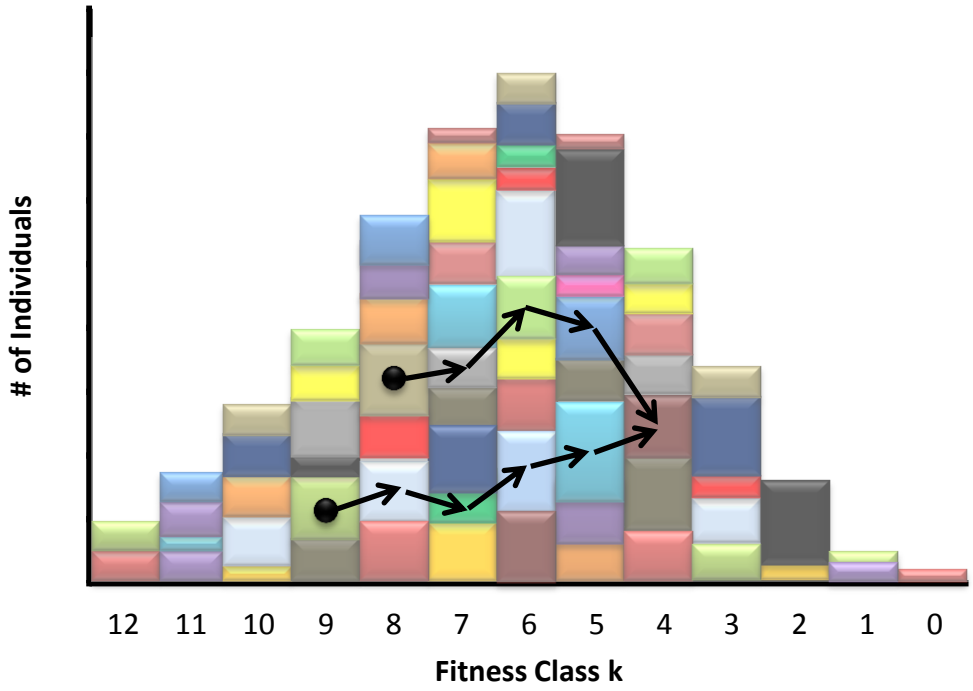

Figure 2.2: **Schematic:** Each fitness class in the population is composed of many lineages, each of which was created by a single mutation and is (in our infinite-sites model) genetically unique. Shown is a schematic cartoon in which each lineage is depicted in a different color. The arrows denote an example of the fitness-class coalescence process for two individuals sampled from classes 8 and 9. These individuals came from different lineages, and these lineages were created by mutations from different lineages within the next most-fit class (as shown by the arrows). The arrows trace the ancestry of the two individuals back through the different lineages that successively founded each other, until they finally coalesce in the class third from right.

interpretation. Fitness class $k-\ell$ has size $n_{k-\ell}$, so the coalescence probability per real generation is $\frac{1}{n_{k-\ell}}$. We will see that each lineage spends of order $s_{k-\ell}$ generations in that class, so the total coalescence probability in this class has the form $\frac{1}{n_{k-\ell}}\frac{1}{s_{k-\ell}}$. This is multiplied by $A_\ell^{k,k'}/2$, which we will show describes the probability that the two individuals are in class $k-\ell$ at the same time. In other words, the probability coalescence occurs in a class equals the inverse population size of the class times

the number of generations lineages spend together in that class. In the following sections of this paper we derive Eq. 2.1 in the two alternative ways mentioned in the Introduction: by explicitly considering the lineage frequency distribution and by following the path summation method of Hudson and Kaplan (1994) and Gordo *et al.* (2002).

2.2.1 Calculating Statistics Describing Sequence Variation

Our approach of treating mutation events as timesteps, and computing coalescence probabilities at each timestep, allows us to make a precise mapping to coalescence theory in which certain quantities have a different meaning than in the traditional theory. In this framework, we can calculate a simple analytic expression for the probability two lineages sampled from particular fitness classes will coalesce in any other fitness class. These fitness-class coalescence probabilities allow us to explicitly calculate the structure of genealogies in this "mutation time." We can then compute the distribution of any statistic describing expected sequence variation by averaging over the fitness classes our original individuals come from. For a statistic x that depends on genealogies between two individuals, for example, we write expressions of the form

$$P(x) = \sum H(k, k')\text{Prob}[k, k' \text{ coalesce in } k - \ell]P(x|k, k', \ell), \tag{2.3}$$

where $H(k, k')$ describes the probability two individuals sampled at random from the population come from classes k and k' respectively.

From the form of these expressions and our simple result for the coalescence probabilities, we can immediately see the main effect of selection on the structure of genealogies. The discussion following Eq. (2.1) implies that the effect of negative selection is similar to that of an effective population size that changes as time recedes into the distant past — i.e. some $N_e(t)$. This intuition has been suggested by earlier work (see e.g. Seger *et al.* (2010)). As we will see, our analysis describes the precise form of $N_e(t)$: it follows the distribution $n_{k-\ell}$ as ℓ increases further to the past, modified by the coefficient $A_\ell^{k,k'}$. We will also see that this picture of time-varying population size has limits: different pairs of individuals have a different $N_e(t)$. As

is clear from Eq. (2.3), these different histories are averaged according to the distribution $H(k, k')$. While it is the average $N_e(t)$ between pairs that determines the distribution of pairwise statistics, this suggests that statistical power may exist in larger samples to distinguish negative selection from neutral population expansion. We explore these general conclusions of our analysis in detail in the Discussion.

Note that in the standard neutral coalescent, one first calculates the distribution of coalescence times and then imagines mutations occurring as a Poisson process throughout the coalescent tree, with rates proportional to branch lengths. In our fitness-class coalescent, by contrast, the coalescence times *are* the mutations. To avoid confusion, from here on we will refer to the effective "generations" in our model as "steps," and refer to the fitness-class coalescent "times" as the "steptimes." We will reserve the word "time" to refer to the actual coalescent time, measured in actual generations.

After determining a fitness-class coalescent tree, we can invert our mapping to determine the structure of genealogies in real time. We will do this by calculating how the steptime in our fitness-class coalescent model translates into an actual time in generations. This will allow us to relate the distribution of branch lengths in steptimes to an actual coalescent tree in generations. We can then treat neutral mutations as is usually done in the standard coalescent: as a Poisson process with probabilities proportional to branch lengths.

Our fitness-time coalescent requires a number of approximations which limit its applicability. Most importantly, we neglect Muller's ratchet, and more generally ignore the effects of fluctuations in the size of each fitness class. We discuss these approximations in more detail below. We find that within a broad and biologically relevant parameter regime they lead to systematic but small corrections to our results. Despite these limitations, our approach also has several advantages relative to previous work. The fitness-time coalescent approach makes many otherwise difficult analytic calculations tractable, allows us to compute the diversity at the selected sites in addition to linked neutral sites, and may offer a useful basis for practical methods of coalescent simulation and inference.

2.3 Model

We imagine a finite haploid population of constant size N. Each haploid genome has a large number of sites, which begin in some ancestral state and mutate at a constant rate. Each mutation is either neutral or confers some fitness disadvantage s (where by convention $s > 0$). We assume an infinite-sites framework, so there is negligible probability that two mutations segregate simultaneously at the same site. We assume that there is no epistasis for fitness, and that each deleterious mutation carries fitness cost s, so that the fitness of an individual with k deleterious mutations is $w_k = (1-s)^k$. Since we assume that $s \ll 1$, we will often approximate w_k by $1-sk$.

The population dynamics are assumed to follow the diffusion limit of the standard Wright-Fisher model. That is, we assume that deleterious mutations occur at a genome-wide rate U_d per individual per generation (with deleterious mutations assumed to be decoupled from selection). We define $\theta_d/2 \equiv NU_d$, the per-genome scaled deleterious mutation rate. Similarly, neutral mutations occur at a rate U_n per individual per generation, and we analogously define $\theta_n/2 \equiv NU_n$. We assume that each newly arising mutation occurs at a site at which there are no other segregating polymorphisms in the population (the infinite-sites assumption).

We focus exclusively on the case of perfect linkage, where we imagine that all the sites we are considering are in an asexual genome or within a short enough distance in a sexual genome that recombination can be entirely neglected. Although our model is defined for haploids, this assumption means that our analysis also applies to diploid populations provided that there is no dominance (i.e. being homozygous for the deleterious mutation carries twice the fitness cost as being heterozygous). In this case, our model is equivalent to that considered by Hudson and Kaplan (1994).

We believe that this is the simplest possible model based on a concrete picture of mutations at individual sites that can describe the effects of a large number of linked negatively selected sites on patterns of genetic variation. It is essentially equivalent to the model described by Charlesworth *et al.* (1993) and Hudson and Kaplan (1994), which has formed the basis for much of the analysis of background selection (Charlesworth *et al.* 1993; Gordo *et al.* 2002; Seger *et al.* 2010).

Our analysis will develop a fitness-class coalescent theory that involves tracing the ancestry of individuals as they change in fitness by acquiring deleterious mutations. In order to do this, we need to first understand the distribution of fitnesses within the population. Since in our model all deleterious mutations have the same fitness cost s, we can classify individuals based on their Hamming class, k, relative to the wildtype (which by definition has $k = 0$). That is, individuals in class k have k deleterious mutations more than the most-fit individuals in the population. Note that not all individuals in class k have the same set of k deleterious mutations. Furthermore, k refers only to the number of *deleterious* mutations an individual has; individuals with the same k can have different numbers of neutral mutations. We normalize fitness such that by definition all individuals in class $k = 0$ have fitness 1. Individuals in class k then have fitness $1 - ks$ (Fig. 2.1).

Haigh (1978) showed that the balance between mutation and selection leads to a steady state in which the fraction of the population in fitness class k, which we call h_k, is given by a Poisson distribution with mean U_d/s,

$$h_k = \frac{e^{-U_d/s}}{k!}\left(\frac{U_d}{s}\right)^k. \tag{2.4}$$

This means that the average fitness in the population is $1 - U_d$, and that $\bar{k} = \frac{U_d}{s}$.

Throughout our analysis, we will assume that the population exists in this steady state mutation-selection balance. In particular, we neglect the fact that in a finite population there will be fluctuations around this h_k. This approximation is central to our approach, and we make it in subtly different ways in both our lineage-structure and our sum of ancestral paths calculations of the fitness-class coalescence probabilities. It will typically be valid in the bulk of the fitness distribution when selection is strong ($Ns \gg 1$); our analysis is limited to this strong selection case and breaks down when $Ns \lesssim 1$. We discuss this approximation in more detail in the Discussion and in Appendix B. We note that this approximation also implies that we assume that Muller's ratchet can be neglected. We will return to the question of the importance of Muller's ratchet in more detail in the Discussion.

We will later need to understand the distributions of timings, $Q_k^{k-1}(t)$, at which an individual mutates from class $k-1$ to class k. We can calculate this by noting that

the probability that an individual in class k arose from a mutation in an individual in class $k-1$ rather than a reproduction event from an individual in class k is

$$\frac{NU_d h_{k-1}}{Nh_k[1-U_d-s(k-\bar{k})]+NU_d h_{k-1}}. \tag{2.5}$$

Substituting in the steady state values for the h_k, and noting that these mutation events are a Poisson process, we find

$$Q_k^{k-1}(t) = ske^{-skt}. \tag{2.6}$$

Note that this calculation is identical to the equivalent distribution of mutation timings computed by Gordo *et al.* (2002) following the approach of Hudson and Kaplan (1994).

2.4 Lineage Structure and the Fitness-Class Coalescence Probabilities

In general, the individuals in a particular fitness class k will not be genetically identical. Rather, there will be a number of different lineages within this class, each lineage created by a deleterious mutation from class $k-1$. We now consider the structure of lineage diversity amongst individuals within a given fitness class in the mutation-selection balance. Note that for our purposes here, we only consider deleterious mutations in defining lineages; we consider the diversity at neutral sites separately below.

Consider a fitness class k, which has an overall frequency h_k (Fig. 2.1b). The frequency h_k is maintained by a stochastic process in which the class is constantly receiving new individuals from class $k-1$ due to deleterious mutations. In our infinite-alleles model, each such mutation creates a lineage which is an allele that is unique within the population. Each lineage fluctuates in frequency for a while before eventually dying out, perhaps after acquiring additional mutations that found new lineages in fitness class $k+1$. At any given moment, there is some frequency distribution of lineages in each class k (see Fig. 2.2). While the identity of these lineages changes over time, there is a probability distribution that at any moment

there is a given frequency distribution of lineages. In steady state, this probability distribution does not change with time.

New lineages are founded in class k at a rate $\theta_k/2$, where

$$\theta_k = 2Nh_{k-1}U_d. \tag{2.7}$$

These individuals are then removed from class k at a per capita rate

$$s_k \equiv -U_d - s(k - \bar{k}). \tag{2.8}$$

We refer to s_k as the *effective selection coefficient* against an allele in class k, because it is the rate at which any particular lineage in class k loses individuals, and we define

$$\gamma_k = Ns_k. \tag{2.9}$$

Using these definitions, we can compute the steady state probability distribution of lineages using the Poisson Random Field model of Sawyer and Hartl (1992). The essential result is that the number of distinct lineages in class k with a frequency between a and b (in the total population) is Poisson distributed with mean $\int_a^b f_k(x)dx$, where

$$f_k(x) = \frac{\theta_k}{x(1-x)} \frac{1 - e^{-2\gamma_k(1-x)}}{1 - e^{-2\gamma_k}}. \tag{2.10}$$

Note that our Poisson Random Field result implies that on average the sum of all the frequencies of all the alleles in fitness class k is simply $h_k = \int_0^1 x f_k(x)dx$, and that the probability that two individuals chosen at the same time at random from fitness class k both come from the same lineage is $\int_0^1 dx x^2 f_k(x)/h_k^2$.

We note that the PRF result involves various implicit approximations, and is valid within a specific parameter regime. Most importantly, we neglect fluctuations in the sizes of each fitness class. This has two main effects. First, it means that we neglect the corresponding fluctuations in the distribution of lineage frequencies $f_k(x)$. Second, it means we are implicitly neglecting the fact that, given a lineage of size x exists in class k, the actual h_k is on average not at its steady state value (e.g. if a high-frequency lineage exists, h_k will tend to be larger). We explain these approximations in detail in Appendix B, and describe an alternative branching process formulation for the lineage structure that corrects for the second effect described above.

2.4.1 The Fitness-class Coalescent Probabilities

We can now calculate the degree of relatedness between two individuals sampled from the population. Our goal is to understand the probability distribution of the fitness-class coalescence steptimes for two individuals chosen at random from the population. We begin by calculating the coalescence probability in each step.

First, imagine that by chance we pick two individuals from the same fitness class k. If the two individuals are from the same lineage, they coalesce within this class. In this case, they are genetically identical and the coalescence steptime is 0. If not, we want to calculate the probability they coalesce in class $k-1$, $P_c^{k,k\to k-1}$. If the lineage of individual A in class k was founded by a mutation from class $k-1$ a time t_1 ago, and the lineage of individual B in class k was founded by a mutation a time t_2 ago, the probability the two individuals came from a common lineage in class $k-1$ is

$$P_c^{k,k\to k-1} = \int dt_1 dt_2 Q_{k,k}^{k-1}(t_1,t_2)\frac{x f_{k-1}(x)}{h_{k-1}}\frac{y}{h_{k-1}}G_{k-1}(y\to x,|t_2-t_1|). \qquad (2.11)$$

Here $Q_{k,k}^{k-1}(t_1,t_2)$ is the joint distribution of t_1 and t_2, x/h_k is the probability one of the individuals came from a lineage of size x given that the lineage exists, $f_k(x)$ is the probability that the lineage exists, and $G_{k-1}(y\to x,|t_2-t_1|)$ is the probability a lineage in class $k-1$ changes in frequency from x to y in time $|t_2-t_1|$ (where y could be 0, corresponding to a lineage that has already mutated back to class $k-2$ by the time the second individual mutates to class $k-1$). The forms of Q and G are described in Appendix A.

If the two individuals coalesced in this first step, the coalescent steptime is 1. If not (which occurs with probability $1-P_c^{k,k\to k-1}$), we have to consider the probability they coalesce at the next step (i.e. in the mutations that took them from class $k-2$ to $k-1$), $P_c^{k,k\to k-2}$, and so on.

So far we have imagined that both individuals that we originally selected from the population came from the same class k. This will not generally be true. Rather, when we pick two individuals at random, they will come from classes k and k' with

probability

$$H(k,k') = \begin{cases} 2h_k h_{k'} & \text{if } k \neq k' \\ h_k^2 & \text{if } k = k' \end{cases} \tag{2.12}$$

For convenience we choose $k \leq k'$. We define $P_c^{k,k' \to k-\ell}$ to be the probability that two individuals from classes k and k' coalesce in class $k-\ell$. Note that $P_c^{k,k' \to k-\ell} = 0$ for $\ell < 0$. For $\ell \geq 0$ we have

$$P_c^{k,k' \to k-\ell} = \int dx dy dt_1 dt_2 Q_{k,k'}^{k-\ell}(t_1,t_2) \frac{x f_{k-\ell}(x)}{h_{k-\ell}} \frac{y G_{k-\ell}(y \to x, |t_2 - t_1|)}{h_{k-\ell}}. \tag{2.13}$$

From the set of coalescence probabilities Eq. (2.13), we can calculate the probability distribution of coalescence steptimes between two individuals. We describe these steptimes by the distribution of classes in which coalescence occurs; given that we pick two individuals from classes k and k' (with $k < k'$ by convention) the probability that they coalesce in class $k-\ell$ is simply

$$\phi_k^{k'}(\ell) = P_c^{k,k' \to k-\ell} \prod_{j=0}^{\ell-1} \left[1 - P_c^{k,k' \to k-j} \right]. \tag{2.14}$$

We note that this expression contains an implicit approximation, as described in Appendix A.

2.4.2 Computing the Coalescence Probabilities

We now have a formal structure describing the structure of coalescent genealogies in the presence of negative selection. It remains, however, to evaluate the coalescent probabilities in each step by evaluating the integrals in Eq. (2.13). We explain the details of this calculation in Appendix A. We find

$$P_c^{k,k' \to k-\ell} = \frac{1}{1 + 2N h_{k-\ell} s(k-\ell)} A_\ell^{k,k'}, \tag{2.15}$$

where $A_\ell^{k,k'}$ is a numerical coefficient which depends on k, k', and ℓ but not on the population parameters,

$$A_\ell^{k,k'} = \frac{\binom{k'}{k-\ell}\binom{k}{k-\ell}}{\binom{k+k'}{2\ell+k'-k}}. \tag{2.16}$$

In Fig. 2.3 we show examples of these coalescence probabilities for different population parameters. We see that the probability of coalescence decreases with increasing selection coefficients and population size.

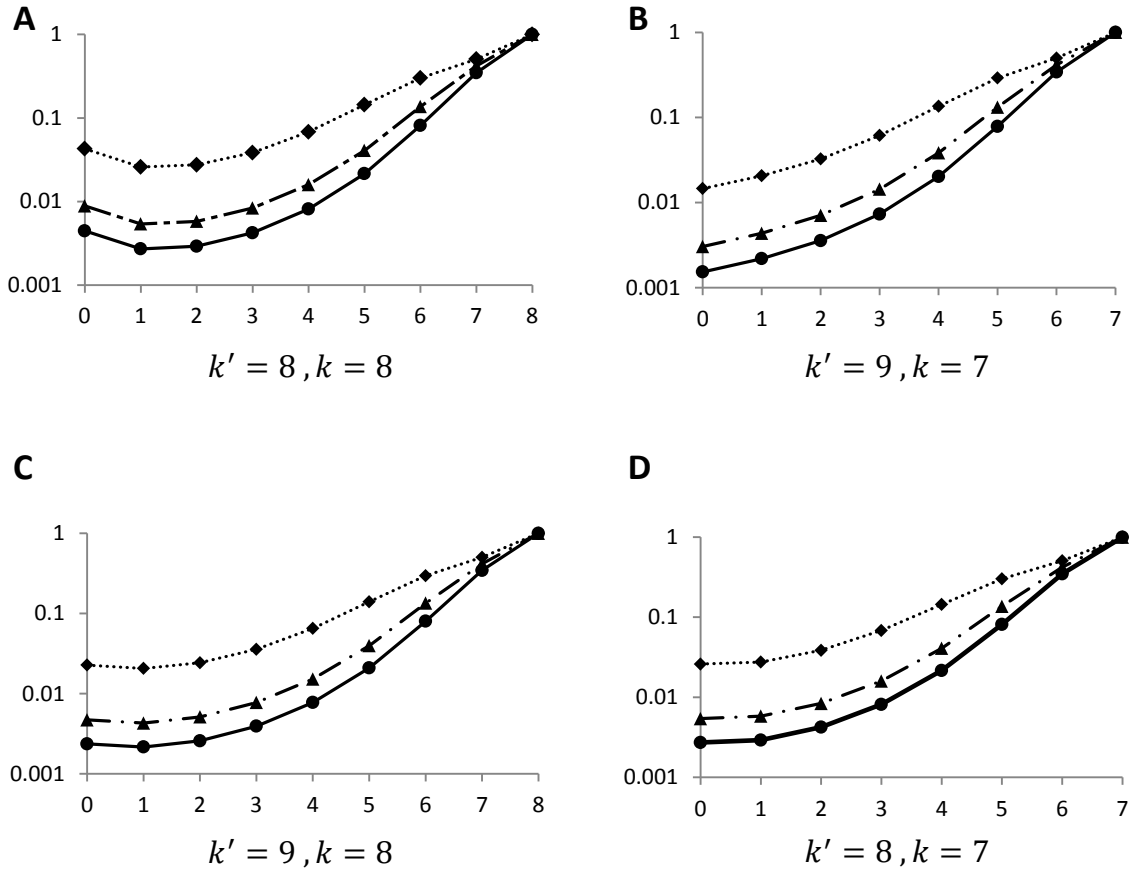

Figure 2.3: **Examples of the Coalescence Probabilities** $P_c^{k,k' \to k-\ell}$**,** for two individuals sampled from fitness classes k and k' to coalesce in class $k - \ell$, shown as a function of ℓ. Here $U_d/s = 8$, $s = 10^{-3}$, and results are shown for $Ns = 10$ (dotted lines), $Ns = 50$ (dashed lines), and $Ns = 100$ (solid lines).

Eq. (2.15) is the complete solution for coalescent probabilities in the non-conditional approximation. This general form for the coalescence probabilities makes intuitive sense. $Nh_{k-\ell}$ is the population size of class $k - \ell$, and $\frac{1}{s(k-\ell)}$ is the average number of

generations that an individual spends in class $k-\ell$ before mutating away. Since the per-generation coalescent probability in a population of size n is proportional to $\frac{1}{n}$, it makes sense that the coalescent probability in class $k-\ell$ is approximately proportional to one over the population size of this class times the number of generations individuals spend in this class. The additional 1 in the denominator captures the fact that the individuals might mutate away from the class before coalescing there (which reduces the average time they spend in the class together). The numerical factor multiplying this basic scaling, $A_\ell^{k,k'}$ comes from the integrals over the probability distribution of mutant timings (i.e. the dt_1 and dt_2 integrals). It reflects the probability that the ancestors of the two individuals we are considering were both in class $k-\ell$ at the same time, since they could not otherwise coalesce there.

From this result, we can also form an intuitive picture of the shape of genealogies in the presence of negative selection. We have just seen that the coalescence probability per actual generation depends on the parameters as $\frac{1}{Nh_{k-\ell}}$, where the relevant value of ℓ increases as we go back in time. Thus the structure of genealogies in the presence of negative selection is similar to having a variable population size as we go back in time. The precise nature of this variable population size is encoded in the fitness distribution $h_{k-\ell}$. For example, if we imagine sampling two individuals from the same below-average fitness class, the probability distribution of their genealogies is like having a population size that initially increases and then decreases as we look backwards in time. Of course, this analogy only goes so far. Most importantly, the coalescent steptimes are related to the statistics describing genetic diversity in a different way from how normal coalescent times are usually related to these statistics. We return to this point in the section on the structure of genealogies below.

2.5 A Sum of Ancestral Paths Approach

We have just computed the fitness-class coalescence probabilities by considering the lineage structure within each fitness class. Kaplan *et al.* (1988) proposed a somewhat different way to look at the same problem: they considered a sample of individuals and, without explicitly describing lineage structure, computed the relative probabil-

ities that the next event to occur backwards in time would involve a mutation or coalescent event. For example, if two individuals are in the same fitness class, the next event could be either coalescence within that class or a mutation event. The rates at which these events occur determines their relative probabilities.

In its original form, this approach used diffusion equations to account for fluctuations in the frequencies of each fitness class h_k. Barton and Etheridge (2004) used this framework to provide a complete solution for the effect of selection at a single site on the structure of genealogies. However, it has not yet proven possible to solve these equations in the more general case of selection at many linked sites. Instead, Hudson and Kaplan (1994) made progress by neglecting fluctuations in the frequencies h_k, the same approximation that is central to our approach. Using this approximation, they derived a recursion relation for the mean time to a common ancestor, their Eq. (12). Gordo *et al.* (2002) used this equation as the basis for a coalescent simulation.

Recursion relations of the Hudson and Kaplan (1994) form can be solved numerically, and have been used to generate data describing coalescent statistics, but have not yet led to an analytic description of the structure of genealogies. We now demonstrate that these numerical methods are equivalent to our lineage-based formalism above, by showing that the Hudson and Kaplan (1994) approach can be used to derive identical analytical formulas for the fitness-class coalescent probabilities. We refer to this as a "sum of ancestral paths" approach, because it relies on summing over all possible paths of individual ancestry through the fitness distribution. The equivalence of this approach to our lineage-structure calculations means that our analytical results in this paper match earlier numerical and simulation results based on the Hudson and Kaplan (1994) formulation.

In order to calculate the coalescence probabilities for a sample of two individuals, we consider the set of all possible ancestral paths these individuals may have followed. Each path is represented by an ordered set of events, backwards in time. These events may either be deleterious mutation events, which move one of the ancestral lineages to the previous fitness class, or coalescence events, which merge the two ancestral lineages. In order for two individuals to coalesce in class $k - \ell$, each ancestral lineage

must undergo a series of deleterious mutation events, bringing them from their initial classes to class $k-\ell$. The lineages must then coalesce before any additional deleterious mutations occur. For example, in order for two individuals sampled from class k to coalesce in class $k-1$, the first event, backwards in time, must be a deleterious mutation. This mutation can occur in either individual. After this event, one of the ancestral lineages is still in class k, while the other is in class $k-1$. The second event, backwards in time, must be a deleterious mutation event in the ancestral lineage that remains in class k. Both ancestral lineages are now in class $k-1$. Finally, the third event must be a coalescent event. Note that there are a total of two paths, since either individual may have been the first to mutate.

The probability of any particular ancestral path is the product of the probability of each event in the path. We saw above that deleterious mutations occur in an individual in class k at rate sk. If the two individuals are in different classes, they are not able to coalesce as the next event. Thus the probability of each possible event is simply:

$$P(\text{1st Event is Del. Mut. in } k|k,k') = \frac{sk}{sk+sk'} \tag{2.17}$$

$$P(\text{1st Event is Del. Mut. in } k'|k,k') = \frac{sk'}{sk+sk'}. \tag{2.18}$$

If the two individuals are in the same class, the next event may either be a coalescent event or a deleterious mutation. Within each class, coalescence is a neutral process that occurs with rate $1/Nh_k$. Therefore, we have

$$P(\text{1st Event is Coal.}|k,k) = \frac{1/(Nh_k)}{sk+sk+1/(Nh_k)} = \frac{1}{1+2Nh_k sk} \tag{2.19}$$

$$P(\text{1st Event is Del. Mut.}|k,k) = \frac{2sk}{sk+sk+1/(Nh_k)} = \frac{2Nh_k sk}{1+2Nh_k sk}. \tag{2.20}$$

These probabilities are analogous to those used by Gordo *et al.* (2002), derived from the framework of Hudson and Kaplan (1994).

Using these probabilities, we can easily calculate the probability of any particular path. In general, in order for two individuals sampled from classes k' and k to coalesce in class $k-\ell$, the ancestral paths must consist of some order of $k'-k+2\ell$ events which include $k'-k+\ell$ deleterious mutation events in the ancestral lineage that began in

k', and ℓ deleterious mutation events in the ancestral lineage that began in k. The path must then conclude with a final coalescent event. Note that there are a total of $\binom{k'-k+2\ell}{l}$ possible paths, reflecting the number of ways to order the mutation events in one lineage with those in the other. To calculate the coalescence probability, we sum the probabilities of each path that results in this particular coalescence event.

We can carry out this sum in the general case by dividing up the $\binom{k'-k+2\ell}{l}$ possible paths according to whether or not the ancestral lineages ever coexisted in each class before class $k-\ell$. Each case leads to a different path probability, and these probabilities can be exactly summed. We carry out this calculation in detail in Appendix A. We find that to leading order in $\frac{1}{1+2Nh_{k-\ell}s(k-\ell)}$, we have

$$P_c^{k,k'\to k-\ell} = \frac{1}{1+2Nh_{k-\ell}s(k-\ell)} A_\ell^{k,k'}, \tag{2.21}$$

which exactly matches our expression for the coalescence probabilities in our PRF approach, Eq. (2.15).

We note that in deriving this result, we have made the same approximations we used in our lineage structure based approach. Thus the results from the PRF method and the sum of ancestral paths are exactly equivalent in the regime where they are valid. However, there are subtle differences in the results to higher orders of the approximations, which provide useful intuition about the process. For example, in the sum of ancestral paths approach it is more natural to calculate $\phi_k^{k'}(\ell)$ directly, without first calculating $P_c^{k,k'\to k-\ell}$, and doing so allows us to compute certain higher-order corrections to the coalescence probabilities. We discuss these details of the correspondence between the approximations used in the two methods in Supplemental Information A.5.

2.6 The Structure of Genealogies and Statistics of Genetic Diversity

We can now use the coalescence probabilities described above to calculate the structure of genealogies in the presence of negative selection. We can then use these genealogies to calculate various statistics describing the genetic diversity within the

population. We know the coalescent probabilities in each step of our fitness-class coalescent process, so in principle we can calculate the probability of any genealogy relating an arbitrary number of individuals using methods analogous to those used in standard neutral coalescent theory. This would then allow us to calculate the distribution of any statistic describing the genetic diversity among these individuals, again using methods analogous to neutral coalescent theory.

Here we will focus on the simplest genealogical relationship: the distribution of the time to the most recent common ancestor of two individuals, which demonstrates the main ideas in the simplest context. This allows us to calculate the distribution of the per-site heterozygosity π. This is the only statistic relevant to a sample of two individuals. In larger samples, the coalescent probabilities between any pair of sampled individuals are independent of those between any other pair that does not share the same most recent common ancestor, so the distribution of per-site heterozygosity we expect within such a sample is closely related to the ensemble distribution of π we calculate here.

In our fitness-class coalescent framework, it is natural to consider diversity at the negatively selected sites separately from diversity at linked neutral sites. We focus first on the distribution of coalescent steptimes and π_d, the per-site heterozygosity at negatively selected sites alone, ignoring neutral mutations. We will then turn to the connection between steptimes and actual times in generations, which will enable us to calculate the distribution of neutral diversity, including the per-site heterozygosity at neutral sites π_n. In analyzing data, we will of course typically not know *a priori* which sites are neutral and which are negatively selected. In such a situation, we merely add up the expected diversity at neutral sites and negatively selected sites, so that the total expected per-site heterozygosity is $\pi = \pi_d + \pi_n$.

2.6.1 Distribution of steptimes and π_d

We begin by imagining that we sample two individuals at random from the *same* fitness class k. If they coalesce in class $k-\ell$, they each acquired ℓ different deleterious mutations to reach class k. Thus the number of negatively selected sites at which

they will be polymorphic is twice their coalescent steptime, $\pi_d = 2\ell$. We therefore have

$$\rho(\pi_d = 2\ell) = \phi_k^k(\ell), \tag{2.22}$$

where $\rho(\pi_d = 2\ell)$ is the probability $\pi_d = 2\ell$.

More generally, if two individuals sampled from classes k and k' coalesce in class $k - \ell$, we have $\pi_d = 2\ell + k' - k$. This means we have

$$\rho(\pi_d = 2\ell + k' - k|k, k') = \phi_k^{k'}(\ell). \tag{2.23}$$

We can average this over the distributions of k and k' to find the distribution of π_d amongst individuals sampled at random from the population. We find

$$\rho(\pi_d) = \sum_{\ell} \sum_{k=0}^{\infty} H(k, k' = k + \pi_d - 2\ell)\phi_k^{k'=k+\pi_d-2\ell}(\ell), \tag{2.24}$$

where the first sum runs from $\ell = 0$ to the largest integer less than or equal to the smaller of k or $\pi_d/2$. Note that in practice we only have to evaluate the sum over k from 0 to a multiple of U_d/s, since $H(k, k')$ will be negligible for larger k.

These results for the distributions of genealogy lengths and of π_d involve several sums. However, all the terms in these sums are straightforward and the numerical evaluations of their values are simple and fast. In Fig. 2.4 we show a representative example of the predicted distribution of the per-site heterozygosity at negatively selected sites, $\rho(\pi_d)$, compared to simulation results. We explore the significance of the shape of the distribution $\rho(\pi_d)$, how this distribution depends on the parameter values, and the source of the small but systematic deviations between the theoretical predictions and the simulation results in the Discussion.

2.6.2 The Relationship between Steptimes and Time in Generations

So far we have focused on the genealogies measured in steptimes, which allowed us to calculate the distribution of heterozygosity among negatively selected sites. We would now like to relate the steptimes to actual times in generations. To do this, we consider the probability that a coalescence event occurred at time t, given two individuals sampled from classes k and k' that coalesced in class $k - \ell$, $\psi(t|k, k', \ell)$.

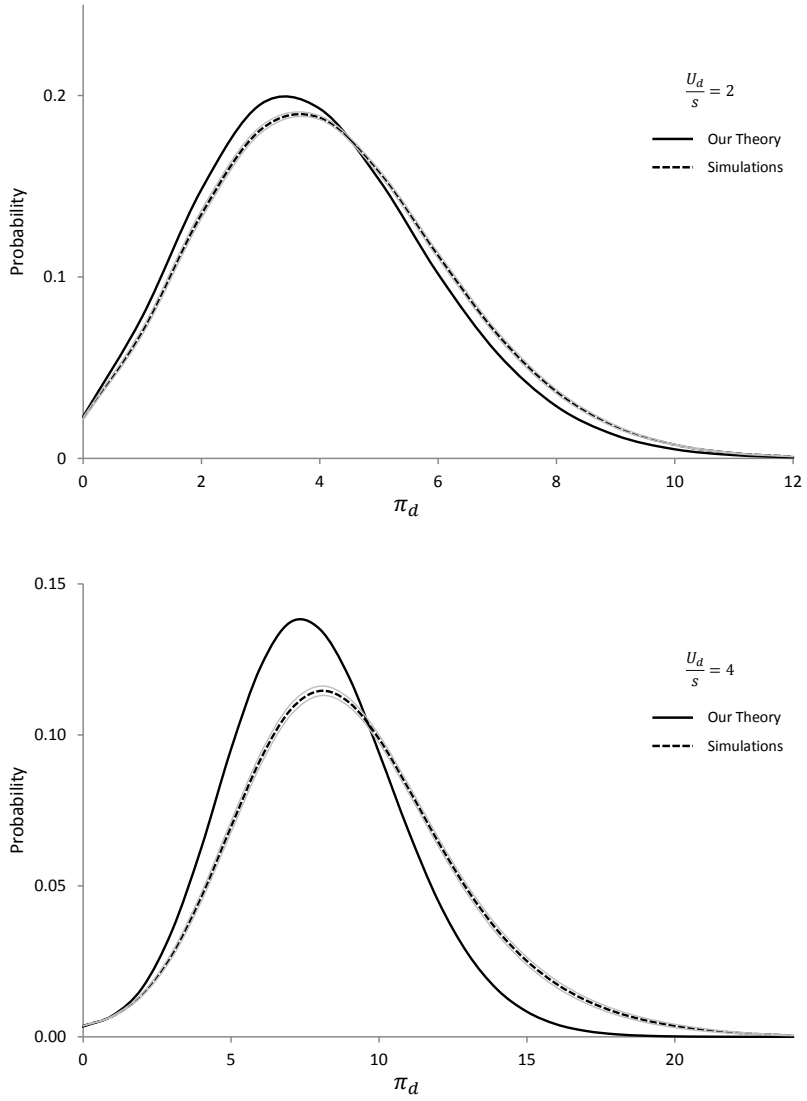

Figure 2.4: **Characteristic Examples of the Distribution of π_d:** Here $N = 5 \times 10^4$, $s = 10^{-3}$ and in **(a)** $U_d/s = 2$, while in **(b)** $U_d/s = 4$. Theoretical predictions are shown as a solid line, simulation results as a dashed line. Simulation results are averaged across at least 300 independent simulations for each parameter set; shaded regions show one standard error in the simulation results. The fit to simulations is good, but we tend to slightly underestimate π_d, and this tendency is worse for larger U_d/s. This is consistent with the effects of Muller's ratchet, which becomes more problematic as we increase U_d/s. This systematic underestimate becomes less severe (for all values of U_d/s) as N increases, as expected, but comprehensive simulations for much larger N are computationally prohibitive.

We compute this distribution in Supplemental Information A.5, and find

$$\psi(t|k',k,\ell) = \sum_{i=0}^{\pi_d-1} s\pi_d(-1)^{\pi_d-i-1}\binom{\pi_d-1}{i}\binom{k'+k}{\pi_d}\frac{B}{A-B}\left(e^{-sBt}-e^{-sAt}\right), \tag{2.25}$$

where we have defined $A \equiv k'+k-i$ and $B \equiv 2(k-\ell)+\frac{1}{Nsh_{k-\ell}}$.

Note that when $Nh_{k-\ell}s(k-\ell) \gg 1$ (the same condition required to neglect fluctuations in h_k, see Appendix B), this expression can be simplified; we find

$$\psi(t|k',k,\ell) = s(\pi_d+1)e^{-s(k'+k)t}(e^{st}-1)^{\pi_d}\binom{k'+k}{\pi_d+1}. \tag{2.26}$$

However, it is important to note that while this approximation may be valid in the bulk of the distribution, it will always fail when coalescence occurs in the zero-class, where $s(k-\ell)=0$. In this case, we must use the more complex expression Eq. (2.25) (or in the case when the coalescence time within the 0-class can be neglected compared to the time taken to descend from the 0-class, the simpler expression described in Eq. (2.39) below).

Averaging over the possible values of k, k', and ℓ, we find the overall distribution of actual coalescent time between two randomly chosen individuals,

$$\psi(t) = \sum_{k'\geq k}\sum_{k=0}^{\infty}\sum_{\ell=0}^{k}\psi(t|k,k',\ell)\phi_k^{k'}(\ell)H(k,k'), \tag{2.27}$$

where the distributions $H(k,k')$, $\phi_k^{k'}(\ell)$, and $\psi(t|k,k',\ell)$ are as given above. However, as we will see below, in calculating neutral diversity we will typically find it easier to work directly with $\psi(t|k,k',\ell)$ rather than this unconditional distribution for $\psi(t)$.

2.6.3 The Neutral Heterozygosity π_n

From the distributions of real times to a common ancestor described above, we can calculate the distribution of π_n, the neutral heterozygosity. Since the neutral mutations occur as a Poisson process with rate U_n, and there are a total of $2t$ generations in which these mutations can occur, π_n follows a Poisson distribution with mean U_nt, where t is drawn from the distribution of coalescence times, Eq. (2.27). We have

$$\rho(\pi_n) = \int_0^\infty \frac{[2U_nt]^{\pi_n}}{\pi_n!}e^{-2U_nt}\psi(t)dt. \tag{2.28}$$

In Fig. 2.5, we compare this distribution of neutral heterozygosity to simulations. We find good general agreement to the shape of the distribution, though there are slight systematic errors (consistent with the effects of Muller's ratchet, which we explore further in the Discussion). Note that, like our results for the diversity at negatively selected sites, these results differ dramatically from the exponential distribution a neutral model or effective population size approximation would predict; we describe these comparisons further in the Discussion.

We note that an alternative way to compute neutral heterozygosity is to further extend the sum of ancestral paths approach which we used above to provide an alternative derivation of the coalescence probabilities. In this formulation, we do not make any connection to real times. However, this approach provides an alternative way to compute the distribution of neutral heterozygosity, $\rho(\pi_n)$. We carry out this computation in Supplemental Information A.6, and show that it leads to results identical to our analysis above.

2.6.4 The Total Heterozygosity π

To calculate the distribution of total heterozygosity $\pi = \pi_n + \pi_d$, we must account for the fact that π_d and π_n are not independent: large π_d means a large coalescent steptime and hence makes a large π_n more likely. The distribution of π_d is given by $\rho(\pi_d)$ above. Above we found $\psi(t|k,k',\ell)$, which implies that

$$\rho(\pi_n|k,k',\ell) = \int_0^\infty \frac{[2U_n t]^{\pi_n}}{\pi_n!} e^{-2U_n t} \psi(t|k,k'\ell) dt. \tag{2.29}$$

We can compute this integral; we find

$$\rho(\pi_n|k',k,\ell) = \sum_{i=0}^{\pi_d-1} \pi_d (-1)^{\pi_d-i-1} \binom{\pi_d-1}{i} \binom{k'+k}{\pi_d} \frac{B}{A-B} \left(\frac{(\frac{2U_n}{s})^{\pi_n}}{(\frac{2U_n}{s}+B)^{\pi_n+1}} - \frac{(\frac{2U_n}{s})^{\pi_n}}{(\frac{2U_n}{s}+A)^{\pi_n+1}} \right). \tag{2.30}$$

Since $\pi_d = 2\ell + k - k'$, this implies

$$\rho(\pi_n|\pi_d) = \sum_{\pi_d = k'-k+2\ell} \rho(\pi_n|k,k',\ell). \tag{2.31}$$

This describes the joint distribution of selected and neutral variation, which is of interest in situations where we know in advance which sites are likely to be neutral

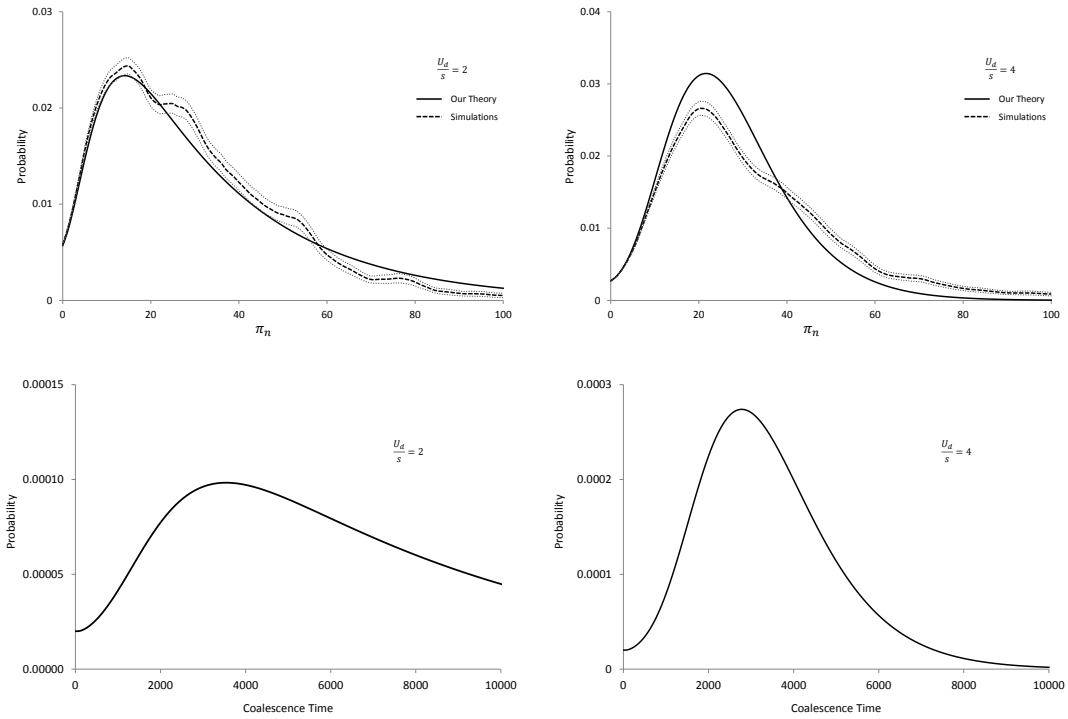

Figure 2.5: **Characteristic Examples of the Distributions of π_n and the Real Coalescent Times: (a)** Theoretical predictions for the distribution of π_n for $U_d/s = 2$, compared to simulation results. **(b)** Theoretical predictions for the distribution of π_n for $U_d/s = 4$, compared to simulation results. Simulation results are averaged across at least 300 independent simulations for each parameter set; shaded regions show one standard error in the simulation results. **(c)** Theoretical predictions for the distribution of real coalescence times for $U_d/s = 2$; note these simply mirror the distribution of π_n, as expected. **(d)** Theoretical predictions for the distribution of real coalescence times for $U_d/s = 4$. In all panels we have $N = 5 \times 10^4$ and $s = 10^{-3}$. Our theory agrees well with the simulations, but note that, as with π_d, we tend to systematically underestimate π_n, and this tendency is worse for larger U_d/s. This is consistent with Muller's ratchet, and as expected becomes more problematic for larger U_d/s. This systematic underestimate becomes less severe (for all values of U_d/s) as we increase N, as expected, but comprehensive simulations for much larger N are computationally prohibitive.

and which are selected (e.g. when analyzing the joint distribution of synonymous and non-synonymous variation). It implies a particular relationship between the observed diversity at selected sites and the reduction in linked neutral variation.

In many situations, however, we will not know which alleles are selected and which are neutral. In this case, we want to understand the distribution of total heterozygosity π, which is given by

$$\rho(\pi) = \sum_{\pi_n+\pi_d=\pi} \rho(\pi_d)\rho(\pi_n|\pi_d). \tag{2.32}$$

This is no more difficult to calculate than $\rho(\pi_n)$, since it involves analogous sums. In Fig. 2.6, we compare this predicted distribution of total heterozygosity to simulations. As with the other aspects of heterozygosity, we find good general agreement to the simulations, with the slight systematic errors that are consistent with the effects of Muller's ratchet.

2.6.5 The Mean Pairwise Heterozygosity

Above we have calculated the distribution of heterozygosity for both neutral and deleterious mutations, as well as total heterozygosity. It is straightforward to average these results to calculate the mean pairwise heterozygosity for both neutral and deleterious mutations; the mean total pairwise heterozygosity is simply the sum of these. In Fig. 2.7 and Fig. 2.8 we show how this mean heterozygosity depends on population size, mutation rate, and selection strength, for neutral and deleterious mutations respectively. We see that the dependence of $\langle\pi_d\rangle$ on the population size is fairly weak. While it increases roughly linearly with N in the weak selection regime, this quickly saturates and for Ns substantially greater than 1 the mean heterozygosity becomes almost independent of population size. The dependence on U_d/s, by contrast, is much stronger. The dependence of $\langle\pi_n\rangle$ on the parameters is also interesting: this depends weakly on the parameters for small N or U_d/s, but for larger N becomes roughly linear. These results make intuitive sense, particularly in light of the "mutation-time" approximation that we introduce in the Discussion, where we discuss these figures in more detail.

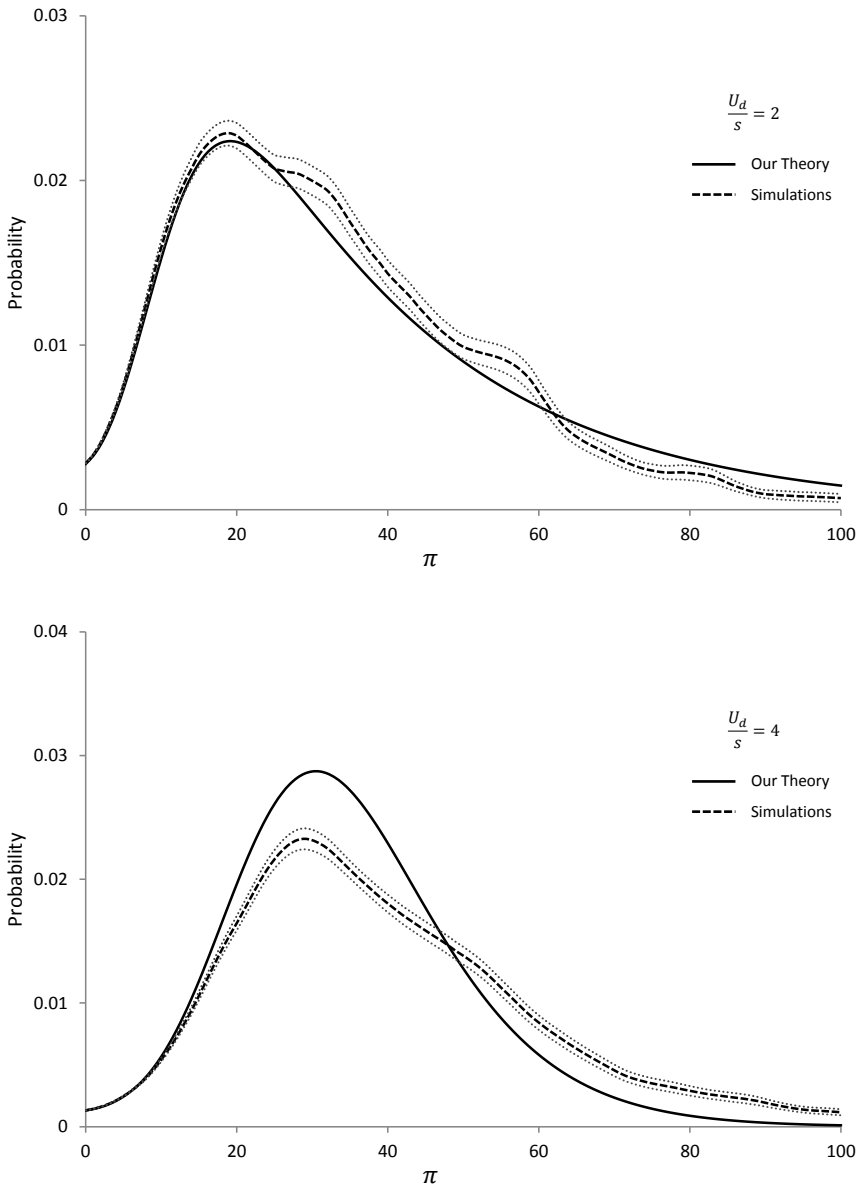

Figure 2.6: **Characteristic Examples of the Distribution of Total Heterozygosity π:** Here $N = 5 \times 10^4$, $s = 10^{-3}$ and in **(a)** $U_d/s = 2$, while in **(b)** $U_d/s = 4$. Theoretical predictions are shown as a solid line, simulation results as a dashed line. Simulation results are averaged across at least 300 independent simulations for each parameter set; shaded regions show one standard error in the simulation results. The fit to simulations is good, but we tend to slightly underestimate π, and this tendency is worse for larger U_d/s. This is for the same reasons as in the distributions of π_n and π_d.

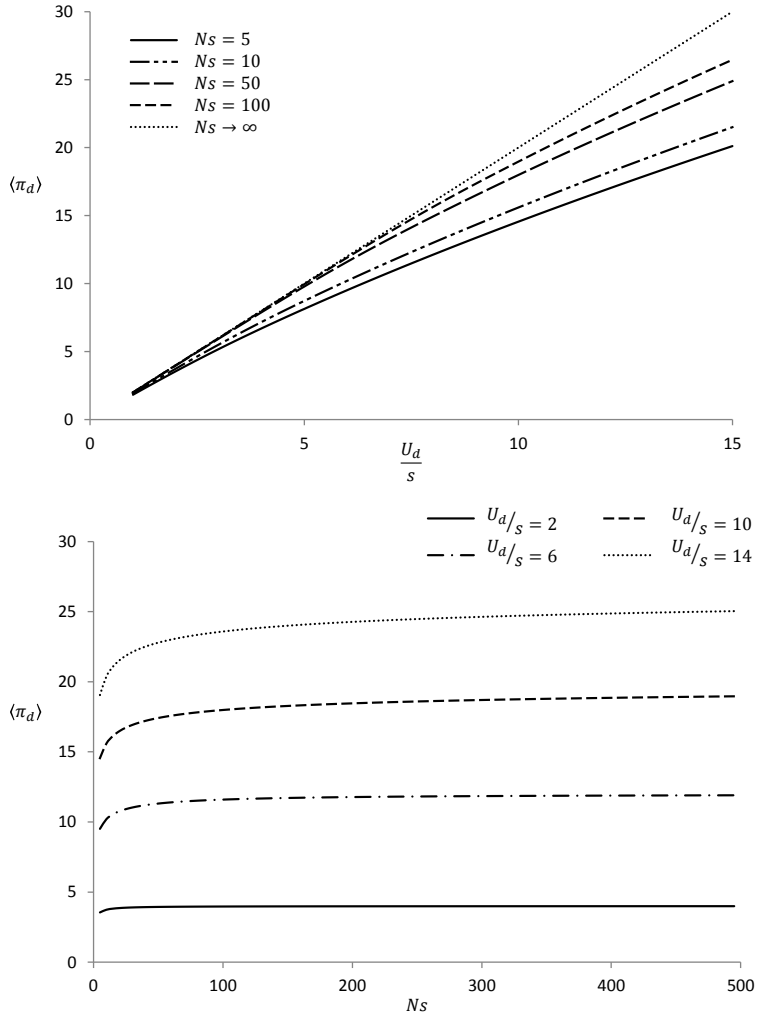

Figure 2.7: **Theoretical Predictions for the Mean Pairwise Heterozygosity at Negatively Selected Sites,** $\langle\pi_d\rangle$, as a function of the parameters. **(a)** $\langle\pi_d\rangle$ as a function of U_d/s for several values of Ns. In the "mutation-time" approximation we expect this to be linear with a slope of 2, since on average individuals are sampled from the mean class at $k = U_d/s$ and coalesce in the 0-class, and hence have $\pi_d = 2U_d/s$. We see that as expected this approximation becomes more and more accurate as Ns increases. For smaller N, there is substantial probability of coalescence in the bulk of the fitness distribution, which is greater for larger U_d/s. Thus the slope of $\langle\pi_d\rangle$ as a function of U_d/s decreases as Ns decreases, and has a downwards curvature. **(b)** $\langle\pi_d\rangle$ as a function of Ns for several values of U_d/s. We see that as Ns becomes large, $\langle\pi_d\rangle$ approaches $2U_d/s$, again consistent with the mutation-time approximation. As Ns decreases, coalescence within the bulk of the fitness distribution becomes more likely, and hence $\langle\pi_d\rangle$ decreases.

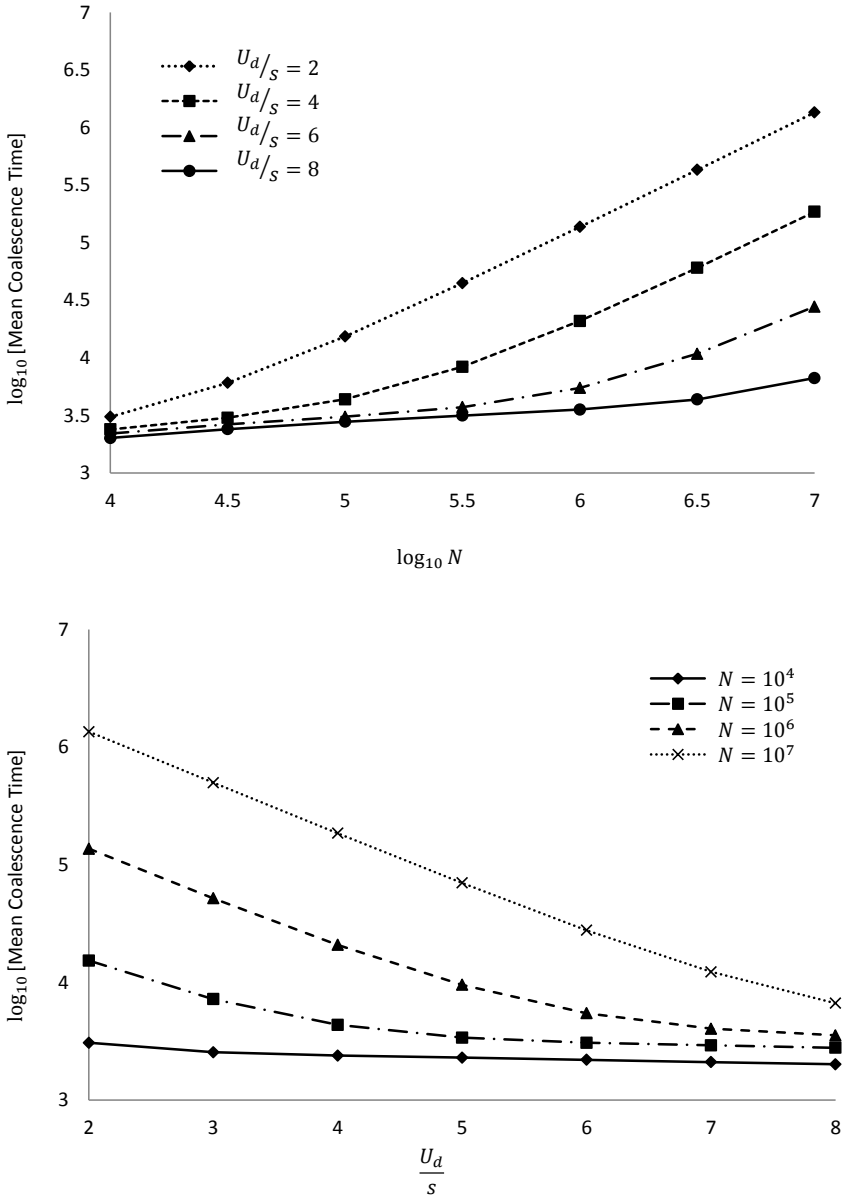

Figure 2.8: **Theoretical Predictions for the Mean Real Coalescence Time $\langle t \rangle$:** In this figure we fix $s = 10^{-3}$ and show the dependence of the mean pairwise heterozygosity on N and on U_d/s. The mean pairwise heterozygosity at neutral sites, $\langle \pi_n \rangle$ is simply $\langle \pi_n \rangle = 2U_n \langle t \rangle$. **(a)** Mean coalescence time as a function of N for various values of U_d/s. We see that $\langle t \rangle$ increases slowly with N until for large enough N the EPS approximation applies and $\langle t \rangle$ becomes linear in N. **(b)** Mean coalescence time as a function of U_d/s for several values of N. For large N, the dependence is roughly linear, consistent with the EPS approximation. For smaller N, coalescence can occur in the bulk of the fitness distribution, reducing the mean coalescence time.

2.6.6 Statistics in Larger Samples

The distributions of π_n and π_d described above are very different from the distributions of heterozygosity expected in the absence of selection. We could certainly measure the distribution of pairwise heterozygosity from a sample of many individuals from a population, and use this to infer the action of selection. However, it may also be useful to understand the expected distribution of other statistics describing the variation in larger samples. One statistic often used to describe variation in larger samples is the total number of segregating sites among a sample of n individuals, S_n. Here we describe how our framework allows us to calculate the distribution of S_3; similar methods can be used to calculate the distribution of S_n for larger n. As we will see, it is unwieldy to calculate closed form expressions for these quantities in our framework, so here we merely lay out a prescription for calculating S_3.

We first consider the distribution of S_3^d, the number of segregating negatively selected sites among three randomly sampled individuals. In order to calculate the probability a sample has a particular S_3^d, we imagine picking three individuals at random from the population and calculate the probability of the coalescence events that lead to that S_3^d. We illustrate such a situation where three individuals are sampled from classes k, k', and k'' in Fig. 2.9. Two of these three lineages coalesced in class k_1. We call the steptime at which two of the three lineages coalesced τ_3 (see Fig. 2.9). We next need to calculate the distribution of τ_2, the total steptime to common ancestry of the three individuals. This time of course cannot be smaller than τ_3. Given values of τ_3 and τ_2, it is clear from Fig. 2.9 that the total number of segregating negatively selected sites is $S_3^d = 2\tau_2 + \tau_3 - (k'' - k) - (k'' - k')$..

Calculating the joint distribution of τ_2 and τ_3 is tedious, because we must sum over all possible orderings of the coalescence events, but it can be computed using either our lineage structure method or the sum of ancestral paths approach. The basic result is analogous to our results for the coalescence steptime between a pair of individuals: coalescence probabilities within a given class are proportional to the inverse size of that class times the number of real generations the ancestors of given individuals typically spend in that class, times a factor that reflects the time that the

ancestors of sampled individuals are present in each class at the same time.

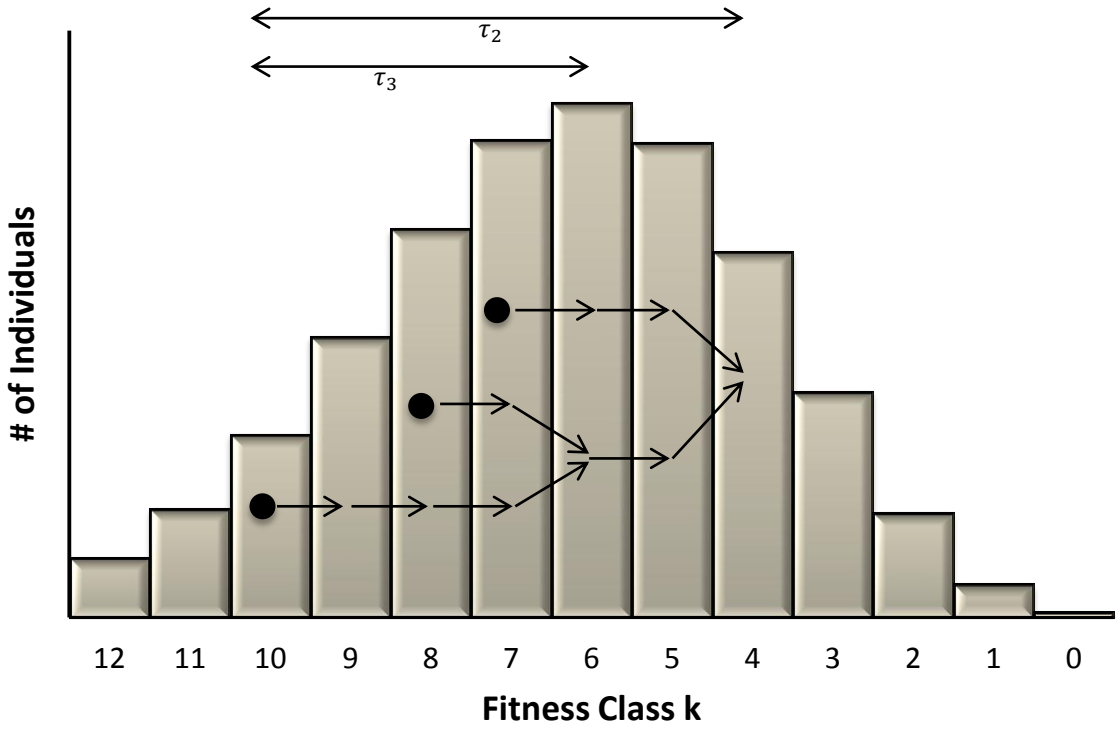

Figure 2.9: **The Fitness-Class Coalescence Process for Three Individuals,** A, B and C, where A and B coalesced τ_3 steptimes ago and C coalesced with the other two τ_2 steptimes ago.

Given a particular value of S_3^d, there is a relationship between the steptimes and actual times (analogous to Eq. (2.25)), which we could use to find the distribution of the total number of segregating neutral sites S_3^n. More complex statistics involving even larger samples can be computed using similar methods.

However, while this analysis provides a prescription for calculating the distribution of S_3^d and S_3^n, it is clear that the full distributions are opaque. In the Discussion we provide a simple approximation for S_n in a specific parameter regime we refer to as the "mutation-time" regime, but the complexities of the general calculation are tangential to the ideas behind our framework, so we do not pursue them further here.

However, these issues will be important to explore in future work aiming to use this framework for data analysis, and our approach here can be used as the basis for genealogical simulations. Further, since our methods allow us to quickly compute the probability of a given genealogical history and to draw a particular genealogy from the appropriate distribution, they may provide a useful basis for importance sampling or MCMC methods to infer selection pressures from data.

2.7 Numerical Simulations of the Genetic Diversity

We compare the predictions of our fitness-class coalescence analysis to Monte Carlo simulations of the Wright-Fisher model. In our simulations, we consider a population of constant size N and we keep track of the frequencies of all genotypes over successive, discrete generations. In each generation, N individuals are sampled with replacement from the preceding generation, according to the standard Wright-Fisher multinomial sampling procedure (Ewens 2004) in which the chance of sampling an individual is determined by its fitness relative to the population mean fitness.

In our simulations, each genotype is characterized by the set of sites at which it harbors deleterious mutations and the set of sites at which it harbors neutral mutations. In each generation, a Poisson number of deleterious mutations are introduced, with mean NU_d, and a Poisson number of neutral mutations are introduced, with mean NU_n; each new mutation is ascribed to a novel site, indexed by a random number. The mutations are distributed randomly and independently among the individuals in the population (so that a single individual might receive multiple mutations in a given generation). The simulations record the time (in generations) at which each distinct genotype was first introduced.

Starting from a monomorphic population, all simulations were run for at least $\frac{1}{s}\ln(U_d/s)$ or N generations (whichever was larger), to ensure relaxation both to the steady-state mutation-selection equilibrium and to the PRF equilibrium of allelic frequencies within each fitness class. The final state of the population — i.e. the frequencies of all surviving genotypes — was recorded at the last generation. In order to produce the empirical distributions of π_d, and π_n shown in Fig. 2.4 and Fig. 2.5,

we averaged across at least 300 independent populations for each parameter set.

Our simulations allow for random fluctuations in the frequencies of each fitness class, and for Muller's ratchet. In most of the parameter regimes we explored, the ratchet proceeded during the simulation, so that the least loaded class at the end of each simulation typically contained anywhere from no deleterious mutations (typical for $U_d/s = 2$) to of order ten (typical for $U_d/s = 4$). We see that despite these effects, our theory agrees well with the simulations, although there are small systematic errors that are consistent with effects of the ratchet. Generally speaking these errors increase as we increase U_d/s, but become less severe for larger N or s. We consider these effects of Muller's ratchet in more detail in the Discussion.

2.8 Discussion

In recent years, both experimental studies and sequence data have pointed to the general importance of selective forces among many linked variants in microbial and viral populations, and on short distance scales in the genomes of sexual organisms (Hahn 2008). Our analysis provides a framework for understanding how one particular type of selection — pervasive purifying (i.e. negative) selection against deleterious mutations — affects the structure of genetic variation at the negatively selected sites themselves and at linked neutral loci. This type of selection is presumably widespread in many populations, in which there is a selective pressure to maintain existing genotypes and mutations away from these genotypes at a variety of loci are deleterious.

A variety of earlier work has addressed aspects of this problem, as described in the Introduction. The key insight of our approach is that instead of following the true ancestral process, we develop a *fitness-class* genealogical approach which focuses on how individuals "move" through the fitness distribution. Here each mutation plays the role of a reproductive event that moves individuals through the fitness distribution, and each fitness class is a "generation" in which coalescence can occur with some probability. We calculate this probability using a simple approximation based on the PRF model of Sawyer and Hartl (1992), rather than by considering the actual reproductive process within that class. By extending formulas originally computed by

Hudson and Kaplan (1994), we showed that these coalescent probabilities can also be computed using a summation of ancestral paths based on the structured coalescent described by Kaplan *et al.* (1988). Hence the conclusions from our analysis also describe the simulations of Gordo *et al.* (2002) and are consistent with all other results based on this structured coalescent approach. Our work is also closely related to recent work in a continuous-fitness model by O'Fallon *et al.* (2010), which uses a similar framework to analyze the weak-selection regime but not the $Ns \gg 1$ situation we study here. We explore the relationship between our analysis and earlier work in more detail in Appendix C.

Our approach leads to simple expressions for the coalescent probability at each step in our fitness-class genealogical process. This makes it a complete effective coalescent theory: using these probabilities, we can calculate the probability that a sample of individuals has any particular ancestral relationship. Our coalescent probabilities are different from those in the standard Kingman coalescent (Kingman 1982), so the structure of genealogies has a different form.

Of course, since our process is an effective rather than an actual coalescent, the relationship between a fitness-class genealogy and the expected statistics of genetic variation given that genealogy is different than in the standard neutral coalescent. Given a particular genealogy measured in steptimes, the numbers of deleterious mutations *are* the coalescent times, and to calculate the statistics of neutral variation we have to make use of the relationship between steptimes and actual coalescence times. This contrasts with the Kingman coalescent, where numbers of neutral mutations are typically Poisson-distributed variables with means proportional to coalescence times (Wakeley 2009). However, we can account for these differences by starting with the distribution of fitness-class genealogies and then converting these genealogies into actual coalescence times.

In this paper, we have used this fitness-class approach to calculate simple statistics describing genetic variation, in particular the distribution of pairwise heterozygosity. This leads to analytic expressions for the quantities of interest, although these expressions involve sums which are most easily calculated numerically. These are easy to compute, and do not become harder to evaluate in larger populations, and hence are

more efficient to evaluate than either simulations or calculations within the ancestral selection graph.

2.8.1 An Intuitive Picture of the Structure of Genealogies

The most important aspect of our analysis is not the specific results for heterozygosity, which match the conclusions of earlier simulations. Rather, the fitness-class coalescent approach allows us to draw several important general conclusions about how negative selection distorts the structure of genealogies. For two individuals drawn from particular fitness classes, the effect of negative selection is similar to that of an effective population size that changes as time recedes into the past. This is consistent with suggestions from earlier work (e.g. the simulation study of Williamson and Orive (2002) and the work of Seger *et al.* (2010)). However, this is not a population size that decreases in a simple way into the past. Our analysis shows the exact form of this time dependent population size. Further, it is clear from our analysis that this is not the only effect of negative selection on genealogies. There are two key complications. First, the statistics of genetic variation (particularly at the deleterious sites themselves) depend on the structure of genealogies differently in our fitness-class coalescent than in the standard neutral coalescent. Second, the time-varying rate of coalescence between a pair of individuals depends on the fitness classes they were sampled from. In other words, different pairs of individuals have a different time-varying effective population size. This suggests that genetic diversity cannot be represented by a single time-varying effective $N_e(t)$ for the whole population, which means that it may be possible to develop statistical tests to distinguish negative selection from population size. All of these general intuitive conclusions about the structure of genealogies in our fitness-class coalescent are illustrated in Fig. 2.10.

We now pause to make this intuitive picture of the shape of typical genealogies more precise. In general the probability that two individuals will coalesce within class k has the form $P_c \approx \frac{A}{2}\frac{1}{n_k sk}$, where n_k is the population size of that class, sk is the effective selection pressure against individuals within that class, and A is a constant

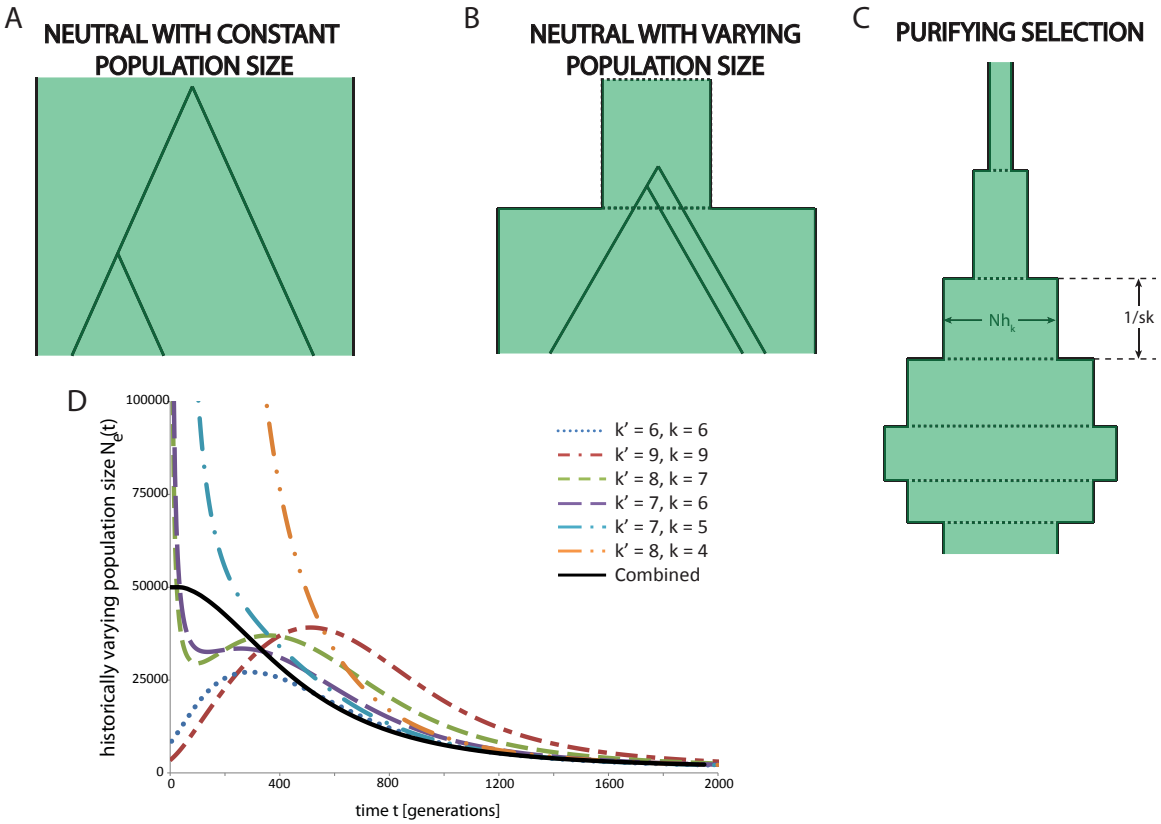

Figure 2.10: **Relationship Between our Results and an Effective Population Size Approximation: (a)** A typical coalescent tree in a neutral population of constant size. The coalescent probability per generation between a random pair of individuals is the inverse population size. Time runs from the past at the top to the present at the bottom. **(b)** An example of a neutral coalescent tree in a population which was smaller in the past than the present. The population size is shown as the width in green. Coalescence events are more likely to occur when the population size is smaller. **(c)** The effective population size history for an individual experiencing purifying selection according to our model. The individual spends on average $\frac{1}{sk}$ generations in class k, which has a total size Nh_k. Note that pairs of individuals are sampled from different classes k (i.e. they are not all sampled from the bottom of this picture). Further, the coalescence probabilities also include a factor of $A/2$, which reflects the probability that two lineages are in the same class at the same time. **(d)** The historically varying effective population size $N_e(t)$ for a pair of individuals sampled from classes k and k', as defined in the text, for several values of k and k'. The $N_e(t)$ for two individuals sampled at random from the whole population is also shown. Here $N = 5 \times 10^4$, $U_d/s = 6$, and $s = 10^{-3}$.

that depends on which classes the lineages began in, but not on any of the population parameters. We have seen that each lineage spends on average $\frac{1}{sk}$ generations in class k. Thus we can think of each individual as seeing a historical effective population size as shown in Fig. 2.10c: it starts in some class k with size n_k and spends $\frac{1}{sk}$ generations in that class before moving to class $k-1$, and so on.

If we sample two individuals, however, they will not always be in the same class at the same time. This effect reduces the coalescence probabilities in each class, as captured by the factor $A/2$. This factor is the average fraction of the $\frac{1}{sk}$ generations each lineage spends in class k that the two lineages spend there together. Alternatively, we can think of this factor as consisting of two parts: A is the probability that the two lineages are ever in the same class at the same time, and $\frac{1}{2sk}$ is the average amount of time that they coexist in the class if they coexist at all (they each spend on average $\frac{1}{sk}$ generations there, but on average overlap for only half this time if they overlap at all). While the two lineages are in the class at the same time, the per-generation coalescent probability is $\frac{1}{n_k}$.

This logic implies that genealogies in the presence of purifying selection look like neutral genealogies with a specific type of historical population size dependence. Imagine for example we picked two individuals from the same fitness class k. They each spend on average $\frac{1}{sk}$ generations in class k, and during that time they have a probability $\frac{A}{2}\frac{1}{n_k}$ per (real) generation of coalescing (this probability includes the fact that on average they are both in the class simultaneously for only a fraction of the mean time each spends there). So roughly speaking, they have an effective population size of $N_e \sim 2n_k/A^{k,k}_{\ell=0}$ for the first $\frac{1}{sk}$ generations. If they fail to coalesce, they then move to class $k-1$, where they spend $\frac{1}{s(k-1)}$ generations and have a probability $\frac{A}{2}\frac{1}{n_{k-1}}$ per generation of coalescing, and hence an effective population size $N_e \sim 2n_{k-1}/A^{k,k}_{\ell=1}$ for this time. If they again fail to coalesce, they move to class $k-2$, and so on.

So far, this picture of a time-dependent population size is rather crude, but we can make it more precise. Specifically, we can write the coalescence probability between

two individuals sampled from class k and k' as a function of time in generations as

$$\psi(t|k,k') = \sum_{\ell=0}^{k} \phi_k^{k'}(\ell)\psi(t|k,k',\ell). \tag{2.33}$$

We can then define the time-dependent effective population size between these individuals, $N_e(t)$, as the inverse probability of coalescence at time t given that coalescence has not yet occurred,

$$\frac{1}{N_e(t)} = \frac{\psi(t|k,k')}{1-\int_0^t \psi(t'|k,k')dt'}. \tag{2.34}$$

In other words, the $N_e(t)$ is defined as usual as the inverse of the probability that the two individuals will coalesce at time t given that they have not yet done so.

We illustrate this precise time-dependent population size $N_e(t)$ in Fig. 2.10d. We see that for two individuals sampled from the same fitness class, $N_e(t)$ typically increases into the recent past and then decreases into the more distant past. This reflects the fact that the two individuals are becoming less likely to be in the same fitness class in the recent past, but that as time recedes into the distant past they are likely to be in the highly fit classes which have smaller n_k. For two individuals sampled from classes near but not identical to each other, $N_e(t)$ starts high and then drops before exhibiting a pattern similar to that among individuals sampled from the same class. This reflects the fact that it takes at least a short time before the two individuals have any chance of being in the same class. Finally, for two individuals sampled from more distant classes, $N_e(t)$ simply declines into the past, both because longer ago they were more likely to be in the same class and more likely to be in the small classes near the high-fitness tail.

Averaging over the whole population, Fig. 2.10d shows the precise time-dependent population size $N_e(t)$ for two randomly sampled individuals. This average $N_e(t)$ initially stays roughly constant as time recedes into the past before decreasing thereafter. For these two randomly sampled individuals, selection is indistinguishable from this particular historically varying population size. The distribution of coalescence times between this pair of individuals looks the same as neutral coalescent histories with this specific population size history. The deleterious mutation rates and selection pressures only matter in that they determine the form of this population size history.

We note that the average $N_e(t)$ shown in Fig. 2.10d implies that recent branches of genealogies will typically be longer relative to ancient branches than we would expect under neutrality. Thus background selection will lead to an excess of low-frequency variants, and hence lead to negative values of Tajima's D, consistent with expectations from previous work (Charlesworth *et al.* 1995; Fu 1997; Gordo *et al.* 2002).

However, a key difference from a neutral population of time-varying size is that, as is clear in Fig. 2.10d, pairs of individuals do not typically come from the same fitness class. Rather, they come at random from different parts of the fitness distribution, and those that come from different places have ancestries characterized by different historically varying population sizes. The total distribution of ancestry is the sum of all of these. In other words, the genetic variation within the population is like that in a population where some individuals had one type of historical population size history, while others had another. If we restrict ourselves to pairwise statistics such as π, the average $N_e(t)$ across pairs of individuals will accurately describe the genetic diversity. However, when we consider appropriately defined statistics in larger samples, the fact that there is no single $N_e(t)$ for the whole population could be important. It remains an interesting question for future work to explore how to exploit this fact to develop statistical tests to distinguish the effects of purifying selection from that of a historically varying effective population size.

2.8.2 Approximations Underlying our Approach

Our analysis relies on several key approximations. First, both our lineage-structure and our sum of ancestral paths methods assume that we can neglect fluctuations in the total frequency h_k of each class. Related to this approximation, we have also implicitly assumed that the probability a lineage in class k reaches a frequency close to h_k can be neglected. In Appendix B, we analyze these approximations in detail and show that they will hold in class k whenever $Nh_k sk \gg 1$. In practice, this condition will often break down in the high and low-fitness tails of the fitness distribution. Fortunately, provided it holds in the bulk of the distribution in which

most individuals will be sampled (which will typically be true provided $Ns \gg 1$), our approach will still be a good approximation. We have also made several other more technical approximations in computing the fitness-class coalescent probabilities. We discuss these in detail in Supplemental Information A.1 and A.4.

Our final and most important approximation is that we assume that Muller's ratchet can be neglected. The ratchet occurs when h_0 fluctuates to 0, so we can think of this approximation as an extreme aspect of neglecting fluctuations in the sizes of each fitness class. This approximation can sometimes be problematic; we discuss it in detail below.

Although we have focused primarily on situations when selection is weak compared to total deleterious mutation rates, our approach is also valid regardless of whether s is strong or weak compared to U_d. However, when selection is sufficiently strong ($Ns \gg 1$ and $U_d/s < 1$), then an effective population size approximation accurately describes the patterns of genetic variation, as we describe below. Thus our methods are primarily useful for situations where selection is weak compared to mutation rates.

2.8.3 Relationship with an Effective Population Size Approximation

Charlesworth *et al.* (1993) considered how selection against many linked deleterious mutations affects linked neutral diversity in a model identical to ours. These authors found that when selection is sufficiently strong, the shape of genealogies and hence the statistics of variation at linked neutral sites is identical to the neutral case, with a reduced effective population size. We refer to this as the effective population size (EPS) approximation.

The idea behind the EPS approximation is that when selection is strong, deleterious mutations are quickly eliminated from the population by selection. Thus if we sample individuals from the population, they must have very recently descended from individuals within the class of individuals which had no deleterious mutations (the 0-class). The EPS approximation assumes that the time for this to happen can be neglected, and that individuals never coalesce before it does. These individuals then coalesce within the 0-class as a neutral process with effective population size equal

to the size of that 0-class, which is $Ne^{-U_d/s}$. Thus the genetic diversity within the population is identical to that in a neutral population of reduced size $N_e = Ne^{-U_d/s}$.

The EPS approximation is valid provided that the neutral coalescence time within the 0-class, t_{neut}, is large compared to the time it takes for a typical individual to have descended from the 0-class, t_{desc}. We know $t_{neut} \sim Ne^{-U_d/s}$, and since a typical individual comes from fitness class $k \sim U_d/s$, we have that $t_{desc} \sim \sum_{j=1}^{U_d/s} \frac{1}{js} \sim \frac{1}{s} \ln\left(\frac{U_d}{s}\right)$. This means that the EPS approximation will be valid provided

$$Nse^{-U_d/s} \gg \ln\left(\frac{U_d}{s}\right). \tag{2.35}$$

Because of the exponential term on the left hand side of this expression, it is clear that the EPS approximation is a strong-selection, weak-mutation limit. It will tend to be valid provided that $Ns > 1$ and $U_d < s$. However, whenever U_d becomes much larger than s, it will typically break down even in enormous populations, as has been suggested by Nordborg *et al.* (1996) and Kaiser and Charlesworth (2009).

Our analysis describes the effects of background selection beyond the EPS approximation. We do not assume that the coalescence time through the fitness distribution is small compared to the coalescence times within the 0-class, or that coalescence cannot occur among individuals carrying deleterious mutations. It is precisely these two effects that lead to distortions away from the neutral expectations, making it impossible to describe genealogies using neutral theory with a revised effective population size. Although our analysis is a generalization of the EPS approximation, it is not inconsistent with it. However, we have focused primarily on situations where the EPS approximation breaks down, and coalescence times through the fitness distribution are large compared to those in the 0-class, because this is the situation where our approach is most useful.

Note also that in many situations it may be the case that there are many linked weakly selected mutations *and* many linked strongly selected mutations. In such circumstances, the process we consider and the EPS approximation can act simultaneously, each for different classes of mutations. Imagine we had one class of mutations with fitness cost s_1 which occur with mutation rate U_1, where $U_1 < s_1$ and $Ns_1 \gg 1$ so that the EPS approximation applies. At the same time, imagine another class of

mutations with fitness cost s_2 which occur with mutation rate U_2, where $U_2 \gg s_2$ so that the EPS approximation breaks down for these mutations. In this case, the genetic diversity we expect to see will be characteristic of our fitness-class coalescent theory (with $U_d = U_2$ and $s = s_2$), but with a reduced effective population size $N_e = Ne^{-U_1/s_1}$. In other words, the strongly selected mutations reduce the effective population size because all individuals are very recently descended from an individual that had no large-effect mutations, but the coalescence time through the distribution of weakly selected mutations cannot be neglected.

2.8.4 A "Mutation-time" Approximation

We have seen that our analysis accounts for two effects missing from the EPS approximation: coalescence events outside the 0-class, and the time it takes for individuals to have descended from the 0-class. Whenever U_d/s and N are both sufficiently large, the former effect can be neglected while the latter is still important, because the number of lineages in each fitness class becomes large and hence coalescence events are very unlikely to occur outside of the 0-class. This leads to an approximation which we can think of as a generalization of the EPS approximation. Rather than considering primarily the diversity generated within the most-fit background, we focus instead on the diversity that accumulates while lineages move between different less-fit backgrounds. Hence we term this approach a "mutation-time approximation" (MTA) for short. In this approximation, we assume that all individuals coalesce within the 0-class, as with the EPS approximation. However, unlike the EPS approximation, we consider the time it took for individuals to descend from the 0-class in addition to the coalescence time within the 0-class. This approximation is valid for large N (when even Nh_1 is enormous compared to $\frac{1}{s}$) so that coalescence always occurs in the 0-class.

In this mutation-time approximation our results become much simpler and provide a useful intuitive picture of the structure of genealogies and genetic variation. Consider the deleterious heterozygosity π_d of two individuals sampled from fitness classes k and k'. In this approximation, these two individuals always coalesce in the

0-class so we always have $\pi_d = k + k'$. Since two individuals are sampled from classes k and k' with probability $H(k, k')$, the distribution of π_d in the population as a whole is extremely simple: we have

$$\rho(\pi_d) = \sum_{k=\pi_d - k'} H(k, k') = e^{-2U_d/s} \frac{1}{\pi_d!} \left(\frac{2U_d}{s}\right)^{\pi_d}. \tag{2.36}$$

This simple approximation makes it clear why the distribution of π_d looks the way it does, and explains how it varies with U_d/s and with N, both in this mutation-time approximation and more generally. For large N, when coalescence outside the 0-class can be neglected, two individuals from class k and k' have $\pi_d = k + k'$. Thus the distribution of π_d has roughly the same shape as the distribution of fitness within the population. The mean π_d is $2U_d/s$, since the average individual comes from class $k = U_d/s$. Smaller and larger π_d are less likely; the distribution of fitness in the population has variance equal to the mean, so the variance of the distribution of π_d is also roughly equal to its mean. As N gets smaller, there is sometimes coalescence outside of the 0-class. This reduces π_d given k and k'. Hence as we reduce N, the distribution of π_d shifts somewhat leftwards, with a peak somewhat below $2U_d/s$, and has slightly more variance relative to the mean since there is a less definite correspondence between k, k', and π_d. Since π_n is determined by π_d, this also explains why the distribution of π_n has the peaked form we observe, and how it depends on U_d/s and N (note that for π_n the coalescence time within the 0-class, which increases linearly with N, must also be included). All of these intuitive expectations are reflected in our results, as shown in Fig. 2.4, Fig. 2.5, Fig. 2.7, and Fig. 2.8. Note for example that in Fig. 2.4, the peak of π_d is slightly below $2U_d/s$ (reflecting the finite population size) and has variance about equal to its mean; we have verified that as N increases the shape of the distribution remains roughly the same, but the mean increases towards $2U_d/s$ and the variance decreases slightly.

More complex statistics of sequence variation are similarly straightforward to calculate in the mutation-time approximation. When considering larger samples, the genetic diversity is determined by the fitness classes these individuals come from, which is always simple since the probability a given individual is sampled from fitness class k is just the Poisson-distributed h_k. This approximation may therefore

prove useful in developing simple and intuitive expressions for various statistics. For example, we can use this approximation to calculate a simple expression for the distribution of the total number of segregating negatively selected sites in a sample of size n, S_n^d, which as we have seen above is otherwise rather involved. We have

$$\rho(S_n^d = x) = \sum_{k_1,k_2,\ldots k_n} h_{k_1} h_{k_2} \ldots h_{k_n}, \tag{2.37}$$

where the sum is over sets of the k_i that sum to x. We find

$$\rho(S_n^d = x) = e^{-nU_d/s} \frac{1}{x!} \left(\frac{nU_d}{s}\right)^x. \tag{2.38}$$

This is a distribution which is peaked around a mean value of $\frac{nU_d}{s}$, for the same reasons the distribution of π_d looks as it does. We note however that as we increase the sample size n the population size N must be even larger for this MTA approximation to hold.

We can also calculate the distributions of actual coalescence times and hence the distributions of statistics describing neutral diversity in the mutation-time approximation. Consider the distribution of the real coalescence time between two individuals chosen from classes k and k'. In the mutation-time approximation where the coalescence time within the 0-class can be neglected, the actual coalescence time is

$$\psi(t|k,k') = s(k+k')e^{-s(k+k')t} \left(e^{st} - 1\right)^{k+k'-1}. \tag{2.39}$$

Averaging over the values of k and k', we have

$$\psi(t) = 2U_d e^{-st - 2(U_d/s)e^{-st}}. \tag{2.40}$$

The distribution of coalescence times once within the 0-class is $\psi_0(t) = \frac{1}{Nh_0} e^{-t/(Nh_0)}$. From this distribution of real coalescence times, we can find the distribution of neutral heterozygosity π_n in the usual way,

$$\rho(\pi_n) = \int_0^\infty \frac{[2U_n t]^{\pi_n}}{\pi_n!} e^{-2U_n t} \psi(t) dt. \tag{2.41}$$

We can immediately see that the average coalescence time in this MTA approximation is $t \approx \sum_0^{2U_d/s} \frac{1}{si} + Nh_0 \approx \frac{1}{s} \ln(2U_d/s) + Nh_0$. We therefore expect that the neutral heterozygosity will on average be

$$\langle \pi_n \rangle \sim \frac{2U_n}{s} \ln\left(\frac{2U_d}{s}\right) + 2Nh_0 U_n. \tag{2.42}$$

The first term in this expression comes from the time to descend through the fitness distribution, while the second term comes from the time to coalesce within the 0-class. If this latter term is large compared to the former, the EPS approximation applies. In the opposite case where the time to descend through the distribution dominates, we can see from the MTA approximation that, as with π_d, the shape of this distribution of π_n is primarily determined by the shape of $H(k, k')$. In this case, the peak in h_k at $k = U_d/s$ leads to a peak in the distribution of real times and hence a peak in the distribution of π_n. The width of the distribution of π_n is somewhat wider, however, since even given individuals coming from fitness classes near the mean, there is a broad distribution of possible real times, and a broad distribution of π_n even given a particular real time.

This average heterozygosity would correspond to an effective population size of

$$N_e \sim \frac{1}{s} \ln\left(\frac{2U_d}{s}\right) + Nh_0, \tag{2.43}$$

but as we have seen this effective population size cannot correctly describe the full distribution of π_n nor its relationship to other statistics describing the genetic diversity. For smaller values of N where the mutation-time approximation breaks down, the average π_n would be somewhat lower than the MTA predicts, and its distribution somewhat broader.

2.8.5 Muller's Ratchet

We have neglected Muller's ratchet throughout our analysis, and assumed that the fitness distribution h_k is fixed. Yet Muller's ratchet will certainly occur, and in some circumstances could have a significant impact on genetic diversity (Charlesworth and Charlesworth 1997; Gordo *et al.* 2002; Seger *et al.* 2010). Thus this is a potentially important omission from our theory. In this section we discuss some of the complications associated with Muller's ratchet that are important to keep in mind when considering our approach. We discuss the parameter regimes where neglecting Muller's ratchet should be reasonable, and those where it is likely to cause more serious problems. We provide rough estimates of how large we expect these problems

to be, and suggest a few possible ways in which future work might incorporate Muller's ratchet into our general framework.

Muller's ratchet causes several related problems within our theoretical framework. First, it causes the values of h_k to change with time, and means they may not always follow a Poisson distribution. This changes the distribution of lineage frequencies within each class, and hence changes the coalescence probabilities. After a "click" of the ratchet, the whole distribution h_k shifts in a complicated way, eventually reaching a new state where it is shifted left (so the class that was originally at frequency h_k is now at frequency h_{k-1}, and so on). In a similarly complex way, the PRF distribution of lineage frequencies in class k shifts from f_k to f_{k-1}, and so on. This naturally changes the coalescence probabilities in each class. Fortunately, since the coalescence probabilities in class k are generally very similar to those in classes $k+1$ or $k-1$, this effect is unlikely to lead to major inaccuracies provided the ratchet does not click many times within a coalescent time. This is true except when we start considering coalescence in classes close to the 0-class, where the k-dependence becomes significant. This can be thought of as an additional problem associated with Muller's ratchet, and is associated with the fact that the ratchet shifts the whole fitness distribution. This effect is easiest to see with an example: imagine we sample two individuals within the k-class, and that these individuals did not coalesce before their ancestors were both in the 0-class. At the time (in the past) when these individuals' ancestors were in the 0-class, this current 0-class might have been the 1-class or 2-class (or higher). Thus these two individuals within the 0-class might not coalesce until, for example, their ancestors were in what is currently the "-2"-class. This clearly means that we might in fact have $\pi_d > 2k$, which our analysis assumes is impossible. In fact, we observe precisely this effect in simulations, and it is the reason why we commonly observe systematic deviations where the simulated values of π_d are larger than our theory predicts.

From this discussion it is clear that the key factor in determining whether Muller's ratchet can reasonably be neglected is how many times the ratchet "clicks" in a coalescence time. We have seen above that an average individual coalesces through

the fitness distribution in a time at most of order $\frac{1}{s}\ln(U_d/s)$ generations. Once within the 0-class, coalescence times are of order $Ne^{-U_d/s}$. We must compare these times to the time it takes for the ratchet to "click." The rate of the ratchet is a complex issue that has been analyzed by Gordo and Charlesworth (2000a), Gordo and Charlesworth (2000b), and Kim and Stephan (2002) in the regime where $Ne^{-U_d/s} > 1$ and by Gessler (1995) in the regime where $Ne^{-U_d/s} < 1$. No general analytic expressions exist which are valid across all parameter regimes. However, provided the ratchet does not typically move a substantial fraction of the width of the fitness distribution in the coalescence time of two random individuals, it will be a small correction to π_d, and neglecting it is a reasonable first approximation. In practice we find in our simulations that for the parameter regimes we consider, π_d is at most of order 2 larger than our theoretical predictions, which would correspond roughly to the effect of a single click of the ratchet during a typical coalescence time.

The discussion above suggests a way to incorporate Muller's ratchet within our theoretical framework, albeit in an ad-hoc way. The ratchet shifts the distribution h_k underneath the fitness-class coalescent process. The details of this shift are complicated, but on average every click of the ratchet shifts the distribution one step to the left. We can define k_{min} to be the number of deleterious mutations (relative to the optimal genotype) in the most-fit individual at any given time. For the case where $Ne^{-U_d/s} > 1$, the rest of the distribution will be approximately a Poisson distribution, but with h_k replaced by $h_{k-k_{\min}}$. Muller's ratchet can then be thought of as a process by which k_{min} increases over time. This increase is a random process, but has some average rate, leading to an average $k_{min}(t)$. As we look backwards in time during the fitness-class coalescent process, the value of k_{min} is decreasing due to Muller's ratchet. This suggests a simple approximation: we replace the actual value of k with an "effective" value of k that accounts for the fact that k_{min} decreases as we look backwards in time. For each step through the fitness distribution, we imagine that k_{min} has decreased by the appropriate amount, and hence the effective value of k in the new fitness class is decreased by less than 1 compared to the old fitness class. When $Ne^{-U_d/s} < 1$ the ratchet is an almost deterministic process, so a similar approximation may prove useful, but in this case the distribution h_k is on average

shifted from the Poisson form (Gessler 1995). To incorporate the ratchet into our analysis in this situation, we first must recalculate the relevant coalescence probabilities given the expected average form of h_k, and then carry out the above program. These and other methods to account for Muller's ratchet remain an interesting topic for future work.

Despite the potential relevance of Muller's ratchet in practical situations, we note that it does not affect our results in the standard coalescent limit. As is apparent from our general expressions for the coalescence probabilities, the structure of our fitness-class coalescent theory does not depend on all three parameters N, U_d, and s independently. Rather, it depends only on the combinations NU_d and Ns. Thus our theory makes sense in the standard limit where NU_d and Ns are held constant while we take $N \to \infty$. In this limit, Muller's ratchet does not occur. Whether this means we can neglect the ratchet for large but finite N depends on the convergence properties of the coalescent limit. This is a difficult limit to explore with simulations, because it requires large population sizes. However, we have used simulations to verify in a few cases that, as expected, increasing N while keeping NU_d and Ns constant does not change the predicted structure of genealogies but decreases some of the systematic differences between theoretical predictions and the simulations which are suggestive of the effect of the ratchet. Note that while this ratchet-free limit does not change the structure of genealogies in our fitness-class coalescent, the distribution of real coalescent times does change, since all real timescales are proportional to s. Thus, as might be expected, we must also take NU_n constant as $N \to \infty$ if we wish neutral diversity to also remain unaffected in this limit.

Note that this ratchet-free limit, while fairly standard in coalescent theory, is somewhat different from the mutation-time approximation we discussed above. Of course, we can easily imagine a population which is large enough that the mutation-time approximation applies, and *then* take the standard coalescent limit.

2.8.6 Conclusion

Our fitness-class coalescent approach provides a framework in which we can compute distributions of genealogical structures in situations where many linked negatively selected sites distort patterns of genetic variation. We have used this framework to calculate the distributions of a few simple statistics describing sequence variation. It remains for future work to use this fitness-class coalescent approach to compute a wide array of statistics to better understand the details of how purifying selection on many linked sites distorts patterns of genetic variation. The eventual goal will be to use our results to help interpret the increasing amounts of sequence data which seem to point to the importance of negative selection on many linked sites.

2.9 Acknowledgments

We thank Daniel Fisher and John Wakeley for many useful discussions, which inspired our fitness-class coalescent approach. MMD acknowledges support from the James S. McDonnell Foundation and the Harvard Milton Fund. AMW thanks the Princeton Center for Theoretical Science at Princeton University, where she was a fellow during some of her work on this paper. LEN is supported by the Department of Defense through the National Defense Science and Engineering Graduate Fellowship Program, and also acknowledges support from an NSF graduate research fellowship. JBP acknowledges support from the James S. McDonnell Foundation, the Alfred P. Sloan Foundation, the David and Lucille Packard Foundation, the Burroughs Wellcome Fund, Defense Advanced Research Projects Agency (HR0011-05-1-0057), and the US National Institute of Allergy and Infectious Diseases (2U54AI057168). Many of the computations in this paper were run on the Odyssey cluster supported by the FAS Sciences Division Research Computing Group at Harvard University.

2.10 Appendix A: The Fitness-class Coalescent Probabilities

2.10.1 PRF Lineage-Structure Approach

In the main text, we used our PRF lineage-structure approach to write an integral expression for the probability $P_c^{k,k'\to k-\ell}$ that two individuals sampled from fitness classes k and k' coalesce in class $k-\ell$, Eq. (2.13) above. In this Appendix, we evaluate this integral to calculate the coalescent probabilities.

Eq. (2.13) depends on the transition probability for the change in the frequency of a lineage from x to y in a time $|t_1 - t_2|$ in class $k-\ell$, $G_{k-\ell}(y \to x, |t_2 - t_1|)$. This transition probability was calculated by Kimura (1955) and can be expressed as an infinite sum of Gegenbauer polynomials. Fortunately, it appears in the context of an integral

$$I_G = \int y G_{k-\ell}(y \to x, |t_2 - t_1|) dy, \tag{2.44}$$

which is simply the average of y over $G_{k-\ell}$. Hence this integral is given by the deterministic result for the change in the frequency of the lineage,

$$I_G = x e^{-s(k-\ell)|t_2-t_1|}. \tag{2.45}$$

Note this deterministic solution simply reflects the exponential decline in frequency of a rare deleterious allele. Substituting Eq. (2.45) into Eq. (2.13), we find

$$P_c^{k,k'\to k-\ell} = \int dx dt_1 dt_2 Q_{k,k'}^{k-\ell}(t_1, t_2) \frac{x^2 f_{k-\ell}(x)}{h_{k-\ell}^2} e^{-s(k-\ell)|t_2-t_1|}. \tag{2.46}$$

The x integral can be evaluated using standard asymptotic methods; we find

$$\int_0^1 dx x^2 f_{k-\ell}(x) \equiv I_x^{k-\ell} = \frac{1}{1 + 2N h_{k-\ell} s(k-\ell)}. \tag{2.47}$$

Note that this and all further expressions for $I_x^{k-\ell}$ incorporate the branching process correction for fluctuations in h_k described in Appendix B. Plugging in this result, we find

$$P_c^{k,k'\to k-\ell} = I_x^{k-\ell} \int dt_1 dt_2 Q_{k,k'}^{k-\ell}(t_1, t_2) e^{-s(k-\ell)|t_2-t_1|}. \tag{2.48}$$

To make further progress, we must understand $Q_{k,k'}^{k-\ell}(t_1, t_2)$, the joint distribution of the times at which individuals sampled from fitness classes k and k' originally

mutated from class $k-\ell$ to class $k-\ell+1$. In general, t_1 and t_2 are not independent, since in order for the two lineages to have coalesced in class $k-\ell$ they must not have coalesced in any earlier classes, which makes them less likely to have been in those classes at the same time. In Supplemental Information A.1, we analyze these distortions and their effects on the coalescence probabilities. Here we make use of a simpler approximation: since the coalescence probability in each step will turn out to be small, conditioning on not coalescing in a particular class does not shift the distribution of mutation timings much. We therefore neglect the complications associated with the probability distributions of the mutant timings conditional on non-coalescence. We refer to this as the non-conditional approximation, and discuss its validity further in Supplemental Information A.1.

In the non-conditional approximation, the times t_1 and t_2 are independent,

$$Q_{k,k'}^{k-\ell}(t_1,t_2) = Q_k^{k-\ell}(t_1)Q_k^{k-\ell}(t_2) \tag{2.49}$$

. We calculate these distributions of mutant timings $Q_k^{k-\ell}(t)$ in Supplemental Information A.2. Plugging these in, and evaluating the integrals as described in Supplemental Information A.3, we find

$$\int dt_1 dt_2 Q_{k,k'}^{k-\ell}(t_1,t_2)e^{-s(k-\ell)|t_2-t_1|} = \frac{\binom{k'}{k-\ell}\binom{k}{k-\ell}}{\binom{k+k'}{2\ell+k'-k}} \equiv A_\ell^{k,k'}. \tag{2.50}$$

Plugging this result into Eq. (2.48), we find $P_c^{k,k'\to k-\ell} = I_x^{k-\ell}A_\ell^{k,k'}$, the result quoted in the main text. We note that $e^{-s(k-\ell)|t_2-t_1|}$ is the probability the ancestor of the first individual to mutate into class $k-\ell$ is still there when the ancestor of the second individual mutated into that class. Thus $A_\ell^{k,k'}$ is the probability that the ancestors of the two individuals were in class $k-\ell$ at the same time, while $I_x^{k-\ell}$ is the probability that they coalesce if so, as described in the main text.

2.10.2 Sum of Ancestral Paths Approach

In the main text, we considered the probability of any particular ancestral path in the history of a sample of two individuals. In this section, we sum over the probabilities of all possible ancestral paths to compute the fitness-class coalescence probabilities.

First, we consider sampling two individuals from the same fitness class k. In order for these two individuals to coalesce in class k, the first event must be a coalescent event. Using the event probabilities computed in the main text, we find $P_c^{k,k\to k} = I_x^k$, equivalent to our earlier lineage-based result. In order for these individuals to coalesce in class $k-1$, the first event must be a deleterious mutation event. Since both individuals' ancestral lineages are currently in class k, the probability the first event is a deleterious mutation event is $1 - I_x^k$. After this event, there is now one ancestral lineage in class $k-1$, and one in class k. The next event must be a deleterious mutation in the latter, which occurs with probability $\frac{k}{2k-1}$. Finally, the third event must be a coalescent event. This implies

$$\phi_k^k(1) = (1 - I_x^k) I_x^{k-1} \frac{k}{2k-1}. \tag{2.51}$$

Note that this logic has given us an expression for the probability that the coalescent steptime is 1, $\phi_k^k(1)$, and not the probability of coalescence in this class given that coalescence has not yet occurred, $P_c^{k,k\to k-\ell}$, because we have already included the probability that the coalescence event does not happen in class ℓ.

We can continue to extend this logic to subsequent fitness classes. For example, for coalescence to occur in class $k-2$, there are six possible paths. We can label them as AABBc, BBAAc, ABABc, ABBAc, BABAc, and BAABc, where A corresponds to a mutation in the first individuals' ancestral lineage, B corresponds to a mutation in the second individuals' ancestral lineage, and c corresponds to a coalescent event. We can calculate the probability of each path. For example,

$$P(AABBc) = \left(\tfrac{1-I_x^k}{2}\right)\left(\tfrac{k-1}{2k-1}\right)\left(\tfrac{k}{2k-2}\right)\left(\tfrac{k-1}{2k-3}\right) I_x^{k-2}. \tag{2.52}$$

The probability of path BBAAc is identical, since it has the same probabilities at each step. However, the remaining four paths have a different probability, because the ancestral lineages exist together in the $k-1$ class at the same time. This distorts the probability of mutations at that step, since coalescence could also have occurred. For paths of this type, we have

$$P(ABABc) = \left(\tfrac{1-I_x^k}{2}\right)\left(\tfrac{k}{2k-1}\right)\left(\tfrac{1-I_x^{k-1}}{2}\right)\left(\tfrac{k-1}{2k-3}\right) I_x^{k-2}. \tag{2.53}$$

We add up each path to find

$$\phi_k^k(2) = I_x^{k-2}\tfrac{k(k-1)}{4(2k-1)(2k-3)}\big(2\big(1-I_x^k\big)+4\big(1-I_x^k\big)\big(1-I_x^{k-1}\big)\big) \quad (2.54)$$

$$= I_x^{k-2}\tfrac{3k(k-1)}{2(2k-1)(2k-3)}\big(1-I_x^k-\tfrac{2}{3}I_x^{k-1}+\tfrac{2}{3}I_x^kI_x^{k-1}\big). \quad (2.55)$$

It is informative to consider the form of this result. The I_x^{k-2} factor is the probability that the two ancestral lineages coalesce in class $k-2$, given that they existed in class $k-2$ at the same time. The remaining factors represent the probability that the two ancestral lineages existed at the same time in class $k-2$. This consists of a leading order term $\frac{k(k-1)}{4(2k-1)(2k-3)}$ (identical to our earlier result for $A_{\ell=2}^k$), multiplied by a correction due to the distortion in paths from the possibility of coalescence in previous steps.

We can continue on to consider the probability of coalescence in class $k-3$. There are now a total of $\binom{6}{3}$ possible paths. These can be split into four types, depending upon whether the two ancestral lineages coexisted in both classes $k-1$ and $k-2$ (e.g. ABABABc), in class $k-1$ only (e.g. ABAABBc), in class $k-2$ only (e.g. AABBABc), or in neither (e.g. AAABBBc). The probability of each type of path is identical, except for a distortion factor $(1-I_x^{k-i})$ for each class $k-i$ in which the two ancestral lineages were together at the same time. The probabilities can be calculated as before, and summed to yield $\phi_k^k(3)$. Using similar logic, we can extend this approach to the situation where two individuals are sampled from different classes, k' and k.

In Supplemental Information A.4, we describe the details of carrying out this summation over all possible paths to determine the coalescent probabilities. We find

$$\phi_k^{k'}(\ell) = I_x^{k-\ell}\frac{\binom{k'}{k-\ell}\binom{k}{k-\ell}}{\binom{k'+k}{k'-k+2\ell}}\left[1-\sum_{i=0}^{\ell-1}\frac{\binom{k'-k+2i}{i}\binom{2\ell-2i}{\ell-i}}{\binom{k'-k+2\ell}{\ell}}I_x^{k-i}+\right. \quad (2.56)$$

$$\left.\sum_{i=0}^{\ell-2}\sum_{j>i}^{\ell-1}\frac{\binom{k'-k+2i}{i}\binom{2j-2i}{j-i}\binom{2\ell-2j}{\ell-j}}{\binom{k'-k+2\ell}{\ell}}I_x^{k-i}I_x^{k-j}-\ldots\right], \quad (2.57)$$

where as always we have assumed $k \le k'$ by convention. The form of this solution is intuitive. The factor $I_x^{k-\ell}$ is the probability of coalescence in class $k-\ell$, given that the two ancestral lineages existed in this class at the same time. The remaining factors reflect the probability that the two lineages are together in class $k-\ell$ at some

point. This consists of a leading order term, which is identical to the $A_\ell^{k,k'}$ calculated previously, times a correction. The correction represents the distortion in the paths due to the possibility that coalescence could have occurred at previous steps. There are a total of $l+1$ terms in the correction, each of which is known and calculable.

Provided that $2Nh_k sk \gg 1$, we can neglect the higher-order terms in Eq. (2.57). This is equivalent to calculating the probability of coalescence in a given class, without considering the possibility that coalescence events could have occurred in previous classes. Thus it converts our expression for $\phi_k^{k'}(\ell)$ into an expression for $P_c^{k,k'\to k-\ell}$. Neglecting these terms also implicitly makes the non-conditional approximation, as we did in the PRF method, because it assumes that the fact that coalescence did not occur in previous classes does not distort the likelihood of taking particular paths. Making this approximation, we find

$$P_c^{k,k'\to k-\ell} = \frac{1}{1+2Nh_{k-\ell}s(k-\ell)} A_\ell^{k,k'}, \tag{2.58}$$

which exactly matches our expression for the coalescence probabilities in the non-conditional approximation in our PRF approach, Eq. (2.15).

The condition $2Nh_k sk \gg 1$ is the condition we are already assuming in treating the frequencies of each class, h_k as constant (see Appendix B). Thus the results from the PRF method and the sum of ancestral paths are exactly equivalent in the regime where they are valid. We discuss the correspondence between approximations in the sum of ancestral paths method as compared to the PRF method in more detail in Supplemental Information A.4.

2.11 Appendix B: Fluctuations in h_k

Throughout our analysis, we have neglected fluctuations in the frequencies of each frequency class h_k. This approximation was necessary to write our PRF expressions for lineage structure, $f_k(x)$, which depend on h_k. Similarly, it was necessary for us to compute the probabilities of each possible ancestral event in our sum of ancestral paths method. In this Appendix, we examine this approximation in detail and analyze its regime of validity.

Fluctuations in the fitness class frequencies affect the coalescence probability within class k in three different ways. First, fluctuations in h_{k-1} affect the rate at which mutations enter class k. When h_{k-1} is larger than average, more mutations occur. Within the PRF method, this means that there will be more small lineages than the steady state $f_k(x)$ accounts for, which reduces the coalescence probability. In the sum of ancestral paths method, this means that the probability of mutation events increases relative to the probability of coalescence events, which similarly reduces the coalescence probability. When h_{k-1} is smaller than average, less mutations occur, and the reverse is true.

Second, fluctuations in h_k affect the coalescence rates within this class. Consider the case where h_k is larger than average. Within the PRF method, this means that the probability that two individuals randomly sampled from class k come from a given lineage of size x is less than our assumption of $\frac{x^2}{h_k^2}$. This reduces the coalescence probability. In the sum of ancestral paths method, this means that the probability of coalescence events decreases relative to mutation events, which similarly reduces the coalescence probability. As before, when h_k is smaller than average, the reverse is true.

The third effect of fluctuations is specific to the PRF method, in which we assumed that the probability two individuals in class k come from a lineage of frequency x (given that the lineage exists) is $\frac{x^2}{h_k^2}$. This implicitly assumes that the fact that there exists a lineage of frequency x in fitness class k does not affect the expected frequency of the class h_k. This is not strictly true: given that there exists a lineage at high frequency, it is likely that h_k is larger than average, and vice versa. In other words, there is a correlation between the size of a lineage and the frequency of the class, so the probability that two individuals picked from a class come from the a lineage of frequency x is not precisely $\frac{x^2}{h_k^2}$. When x is large, this expression overestimates the probability two individuals are from the same lineage, since given that those high-frequency lineages exist, h_k will be larger than average. Similarly (though less dramatically), when x is small our expression underestimates the probability two individuals are from the same lineage.

Note that this third effect of fluctuations is distinct from the second effect above.

The second effect describes fluctuations in h_k that are uncorrelated to the frequency of a particular lineage. It thus applies to both the PRF and sum of ancestral paths methods; it reflects the general fact that when h_k is larger coalescence is less likely. The third effect, on the other hand, reflects the fact that if we assume we sample an individual from a lineage of size x, this biases the value of h_k. Since our sum of ancestral paths method never makes any references to lineages, this third effect of fluctuations only applies to the PRF method.

These three effects all depend on the size of the fluctuations relative to the average size of the each fitness class. Thus neglecting fluctuations will be a good approximation provided that the fluctuations in h_k are small compared to h_k. To determine when this will hold, we note that each lineage in class k can reach, at most, a maximum size of order $\frac{1}{s_k}$ individuals (selection prevents any individual lineage from becoming more common than this). The total number of individuals in the class is on average Nh_k. This means that, provided that $Nh_k \gg \frac{1}{s_k}$, each fitness class is made up of many individual lineages. Thus we would expect that the fluctuations in the sizes of each one would tend to cancel, and the overall fluctuations in h_k should be negligible provided that this condition holds.

To make this intuition more precise, we must calculate the variance in h_k and compare it to h_k. In principle this information is contained in our PRF expressions, but it is much simpler to compute using a continuous-time branching process method. That is, rather than use a diffusion approximation to describe the dynamics of each lineage, we use a continuous-time branching process. As before, we imagine that new lineages in class k are created at a rate $\theta_k/2$. In steady state there will be some time-independent probability that there are n total individuals across all the lineages in the class, $P(n)$. Note that on average we must have $n/N = h_k$, and that $P(n)$ contains information on the fluctuations in the h_k. We first compute the generating function for $P(n)$,

$$H(z) \equiv \sum_{n=0}^{\infty} P(n) z^n. \tag{2.59}$$

To do so, we start by computing the generating function for the probability distribution of the number of individuals from each lineage, as described by Eqs. (7-9) of

Desai and Fisher (2007). We substitute this expression into Eq. (24) of Desai and Fisher (2007) and integrate. We find

$$H(z) \equiv \sum_{n=0}^{\infty} P(n,t) z^n \equiv \langle z^n \rangle = \left[\frac{s}{1 - z(1-s)} \right]^{\frac{\theta}{2(1-s)}}, \tag{2.60}$$

where angle brackets denote expectation values, and we have suppressed the k subscripts. Note that this calculation is based on a continuous-time branching process, in which individuals have a different distribution of offspring number than in a Wright-Fisher process, leading to a transient distribution of the frequencies of individual lineages that is half as large as in the Wright-Fisher model for lineages of substantial frequency. Thus to make comparisons with the Wright-Fisher process, we have to take $\theta \to 2\theta$ (as we would in comparing Wright-Fisher to Moran models), as described by Desai and Fisher (2007).

Eq. (2.60) describes the fluctuations in the size of an individual fitness class: the mean, variance, and higher moments of n can be easily computed by taking derivatives of $H(z)$. Thus we can immediately compute $Var(h_k)/h_k$ using standard generating function methods. We find that in fact the fluctuations in h_k are indeed negligible provided that

$$N h_k s_k \gg 1. \tag{2.61}$$

In practice, this condition will often break down in the high and low-fitness tails of the fitness distribution. Fortunately, provided it holds in the bulk of the distribution in which most individuals will be sampled, which will typically be true provided $Ns \gg 1$, our approach will still be a good approximation.

2.11.1 Correcting for Correlations between the Size of a Lineage and the Frequency of the Fitness Class

All three effects of fluctuations in h_k described above are negligible in the same parameter regime, $N h_k s k \gg 1$. However, the fact that the third effect applies only to our PRF result obscures the precise relationship between our two approaches, and the relationship to earlier work. Further, relaxing this approximation provides a useful comparison of the subtle differences between the assumptions underlying the

approaches. Thus we describe here an alternative approach to understanding the lineage structure in a fitness class which allows us to account for these correlations between the size of a lineage, x, and the frequency of the fitness class, h_k.

We first note that, in his original calculation of the neutral ESF, Ewens (1972) used a diffusion result, $f(x)$, roughly analogous to our PRF expression to describe the probability that there exists a lineage with frequency x in the population at a given time. However, Ewens' $f(x)$ was derived as the solution to the diffusion approximation to the K-allele Wright-Fisher process, in the limit of infinite alleles. This process explicitly imposes the constraint that the sum of all lineages in the population at a given time must add to 1. This means that there is no correlation between the size of a lineage and the total number of individuals in the population.

The PRF calculation of the lineage structure does not involve this explicit constraint. This is what makes it possible to compute a simple analytical expression for $f_k(x)$. This lack of constraint means that the PRF result admits fluctuations in h_k, which lead to corresponding correlations between x and h_k. We could partially avoid this by defining $\gamma_k = Nh_k s_k$, rather than Nh_k, as we have so far. This would effectively mean that each lineage is assumed to be diffusing between 0 and h_k rather than between 0 and 1, and forbid any lineage from reaching a frequency larger than h_k. Thus it reduces the discrepancies associated with the correlations between x and h_k. However, even with this redefinition, there is no constraint that the lineages in a given class all add to precisely h_k, and so correlations still exist.

To correct exactly for the effects of correlations between x and h_k, we extend the continuous-time branching process model introduced above. We now imagine that there are B sites in the genome, each of which can mutate to create a new lineage in class k. In the large-B limit, each distinct lineage in class k arose from a mutation at a different site in the genome (and we will later make the infinite-sites assumption $B \to \infty$, which makes this exactly true). The rate at which new mutations found lineages in class k due to mutations at a specific one of these B sites is $\frac{\theta_k}{2B}$. This means that, analogous to Eq. (2.60), the generating function for the probability that

there are n mutations at a particular site i in class k is

$$H_i(z) = \left[\frac{s}{1-z(1-s)}\right]^{\frac{\theta}{B(1-s)}}, \tag{2.62}$$

where again we have suppressed the k subscripts and we have taken $\theta \to 2\theta$ to match to the Wright-Fisher model as described above.

If we define $n_{i,k}$ to be the total number of mutants at site i in class k, we have that

$$\sigma_k \equiv \sum_{i=1}^{B} n_{i,k} \tag{2.63}$$

is the total number of individuals in the class (note that on average we expect $\sigma_k = Nh_k$). We now imagine that we sample some number m individuals from class k. The probability that they are all from the same lineage is

$$J_m^{(k)} = \left\langle \sum_{i=1}^{B} \frac{n_{i,k}^m}{\sigma_k^m} \right\rangle = \left\langle \frac{n_{1,k}^m}{(n_{1,k}+...n_{1,B})^m} + \frac{n_{2,k}^m}{(n_{1,k}+...n_{1,B})^m} + ... + \frac{n_{B,k}^m}{(n_{1,k}+..n_{1,B})^m} \right\rangle. \tag{2.64}$$

Note this has the same form as our PRF expression, except we are averaging over $\frac{n_i^m}{\sigma^m}$ rather than averaging over n_i^m and *then* dividing by the average σ^m. In other words, we are explicitly accounting for the correlations between x and h_k.

We can rewrite Eq. (2.64) using the identity

$$\frac{1}{\sigma_k^m} = \int_0^\infty \frac{x^{m-1}}{(m-1)!} e^{-x\sigma_k} dx. \tag{2.65}$$

This identity can easily be verified by integrating the RHS by parts. Using this, and noting that lineages at each of the B sites are independent, we find

$$\begin{aligned} J_m^{(k)} &= \left\langle \sum_{i=1}^{B} n_i^m \int_0^\infty \frac{x^{m-1}}{(m-1)!} e^{-x\sigma_k} dx \right\rangle \\ &= B \int_0^\infty \frac{x^{m-1}}{(m-1)!} \langle n_1^m e^{-x\sigma_k} \rangle dx \\ &= B \int_0^\infty \frac{x^{m-1}}{(m-1)!} \langle e^{-xn_i} \rangle^{B-1} \langle n_1^m e^{-xn_1} \rangle dx. \end{aligned} \tag{2.66}$$

The first expectation value inside the integral can be computed by noting that

$$\langle e^{-xn_i} \rangle = H(z = 1-x) = \left[1 + x\frac{1-s}{s}\right]^{\frac{\theta}{B(1-s)}}. \tag{2.67}$$

Differentiating this result m times with respect to x results in an expression for $\langle n_1^m e^{-xn_1} \rangle$. Plugging these results in and integrating, taking the limit $B \to \infty$, and neglecting higher order terms in s, we find

$$J_m^{(k)} = \theta \sum_{j=0}^{m-1} (-1)^j \binom{m-1}{j} \frac{1}{\theta + j} = \frac{(m-1)!}{\prod_{j=1}^{m-1}(\theta + j)} = \frac{1}{\binom{\theta+m-1}{\theta}}. \tag{2.68}$$

If we were to use the original PRF result to calculate the probability two individuals sampled simultaneously from class k are from the same lineage, we would find $\int_0^1 \left(\frac{x}{h_k}\right)^2 f_k(x)dx = \frac{1}{\theta}$. Using our branching process result for $J_2^{(k)}$, we see that correcting the PRF result for the third effect of fluctuations in h_k yields the modified probability $\frac{1}{1+\theta_k}$. As expected, the branching process result precisely matches the sum of ancestral paths approach, which is also unaffected by this third effect of fluctuations in the h_k. All of the formulae quoted in the main text and shown in the figures incorporate this correction, which appropriately handles the correlations between the frequency of an individual lineage and the size of the fitness class.

2.12 Appendix C: Relation to Previous Work

In this Appendix we compare our analysis to related work, and summarize the key approximations that we and others have used. We have presented two main approaches to calculating coalescence probabilities in this paper. The first approach is based on the lineage structure within each fitness class, described using a PRF-based method. The second approach involves summing over all possible ancestral paths, based on the structured coalescent framework introduced by Kaplan *et al.* (1988) and Hudson and Kaplan (1994, 1995b). We show in this paper that both approaches involve closely related approximations and yield equivalent expressions for the coalescence probabilities.

Historically, attempts to describe the coalescent process in the presence of selection go back to the structured coalescent introduced by Kaplan *et al.* (1988). These authors considered a sample of individuals from given fitness classes and computed the relative probabilities that the next event to occur backwards in time would involve a mutation or coalescent event, without explicitly describing lineage structure.

	this work	Hudson & Kaplan 88	Hudson & Kaplan 94,95	Gordo et al 02	Charlesworth et al 93	Barton & Etheridge 04	Seger et al.10	O'Fallon et al. 10
analytical expressions for genealogy structure	x				x	x		x
accounts for frequency class fluctuations (valid for $Ns \sim 1$)		x				x	x	x*
valid for $Nse^{-U/s} << \ln[U/s]$	x	x	x	x		x	x	x
valid for $Ns \gg 1$	x	x	x	x	x	x	x	
valid for many classes	x	x	x	x	x		x	x
accounts for Muller's ratchet		x				x†	x	
discrete fitness classes	x	x	x	x	x	x	x	

Table 2.1: A summary of related approaches to the coalescence process in the presence of purifying selection. *Addresses $Ns \sim 1$ situation, but assumes deterministic fitness distribution. †Within a two-class framework.

In their original work, Kaplan *et al.* (1988) used a full stochastic description of the frequencies of each fitness class, in which one keeps track of the probability distribution of these frequencies to account for selection. They derived diffusion equations for the transition probabilities between states. This approach is very general, but as a result is complex and requires numerical evaluation. Barton and Etheridge (2004) developed this diffusion approach to compute the effect of selection on genealogies in a system in which selection acts only on a single locus.

Hudson and Kaplan (1994) later simplified their original structured coalescent approach to describe the case where fluctuations in the frequencies of fitness classes can be neglected. In this deterministic approximation, they showed that one can compute very simple expressions for the relative probabilities of the next event to occur backwards in time in the history of a sample. In this manner, Hudson and Kaplan (1994) were able to generate a simple recursion relation for the mean time to a common ancestor, their Eq. (12). Gordo *et al.* (2002) used this equation as the basis for a coalescent simulation, and Zeng and Charlesworth (2011) recently extended this method to describe the joint effects of recombination and background selection.

Recursion relations of the Hudson and Kaplan (1994) form can be solved numerically, and have been used to generate data describing coalescent statistics, but have not yet led to an analytic description of the structure of genealogies in the pres-

ence of negative selection at many linked sites. In this paper we have shown that one can sum over ancestral paths within this framework, to derive analytical formulas for the coalescence probabilities which are equivalent to those computed from our lineage-based formalism. This equivalence means that our analytical results in this paper match earlier numerical and simulation results based on the HUDSON and KAPLAN (1994) formulation. However, like the HUDSON and KAPLAN (1994) framework, neither of our approaches in this paper account for fluctuations in the frequencies of fitness classes.

In reality, the frequency of each fitness class will fluctuate due to genetic drift. As we have described in Appendix B, these fluctuations are substantial in classes whose deterministic size is small compared to the inverse of the effective selection pressure against individuals in that class, $Nh_k sk < 1$. This leads to important effects on the structure of genealogies if most fitness classes through the bulk of the fitness distribution fluctuate substantially. This will occur whenever $Ns \lesssim 1$, so fluctuations must therefore be taken into account for small Ns. While the diffusion approach of KAPLAN *et al.* (1988) in principle provides a complete solution to this problem for all values of Ns, this formalism and the related results of BARTON and ETHERIDGE (2004) are computationally strenuous. There remains a need for further work on accurate but more analytically tractable approaches which are able to account for the frequency fluctuations.

We note that the work of O'FALLON *et al.* (2010) and of HERMISSON *et al.* (2002) introduced analytical approaches valid for the case of $Ns \sim 1$, although these methods are not based on a model related to the ideas of KAPLAN *et al.* (1988). We also note that the problem of fluctuating fitness class sizes has been considered in the case of other problems (for example, forward selection COOP and GRIFFITHS (2004)), but a detailed discussion is outside the scope of this work.

Neglecting the fluctuations in fitness class frequencies is in principle reasonable when $Ns \gg 1$. However, we note that even when $Ns \gg 1$, the sizes of the smallest fitness classes near the tails of the distribution may still fluctuate substantially. Muller's ratchet is one aspect of this general effect. Recently SEGER *et al.* (2010) extended the simulation scheme of GORDO *et al.* (2002) to address this problem by

first doing a forward-time simulation, recording the fluctuations in the classes (including Muller's ratchet) from this simulation, and then putting these fluctuations into a backwards simulation by hand. Our methods do not account for these effects. They are therefore less general than the work of Seger *et al.* (2010), and break down due to fluctuation effects more quickly as Ns decreases. On the other hand, our analysis does not rely on forward simulations and is able to compute simple analytic expressions for coalescence probabilities.

We also note that although we consider the large Ns approximation, our approach has a broader range of applicability than the effective population size approximation, which assumes that the coalescence time is dominated by the time to coalescence within the most-fit class. For the EPS approximation to be valid requires that this latter time ($\sim Ne^{-U_d/s}$) is small compared to the time average individuals took to descend from the most-fit class ($\sim \frac{1}{s}\ln Ns$). Thus for the EPS approximation to hold, we require $Ne^{-U_d/s} \gg \frac{1}{s}\ln[U_d/s]$, not just $Ns \gg 1$. Thus we can easily have $Ns \gg 1$, yet $Nse^{-U_d/s} \ll \ln[U_d/s]$, in which case the EPS approximation breaks down and yet our approach is still valid.

Chapter 3

The Structure of Allelic Diversity in the Presence of Purifying Selection

In the absence of selection, the structure of equilibrium allelic diversity is described by the elegant sampling formula of Ewens. This formula has helped shape our expectations of empirical patterns of molecular variation. Along with coalescent theory, it provides statistical techniques for rejecting the null model of neutrality. However, we still do not fully understand the statistics of the allelic diversity expected in the presence of natural selection. Earlier work has described the effects of strongly deleterious mutations linked to many neutral sites, and allelic variation in models where offspring fitness is unrelated to parental fitness, but it has proven difficult to understand allelic diversity in the presence of purifying selection at many linked sites. Here, we study the population genetics of infinitely many perfectly linked sites, some neutral and some deleterious. Our approach is based on studying the lineage structure within each class of individuals of similar fitness in the deleterious mutation-selection balance. Consistent with previous observations, we find that for moderate and weak selection pressures, the patterns of allelic diversity cannot be

described by a neutral model for any choice of the effective population site. We compute precisely how purifying selection at many linked sites distorts the patterns of allelic diversity, by developing expressions for the likelihood of any configuration of allelic types in a sample analogous to the Ewens sampling formula.

3.1 Introduction

In any evolving population, new clonal lineages are constantly being created and destroyed. The balance between the creation of lineages by new mutations and their destruction by natural selection and genetic drift determines the statistics of the clonal structure of the population. In the absence of natural selection, Ewens (1972) computed an elegant sampling formula describing the clonal structure of a neutral population, and explained how the allelic (i.e. lineage) configuration in a sample of individuals from the population provides a window into this clonal structure.

Natural selection distorts the clonal structure of a population away from this neutral expectation. Of particular interest is purifying (negative) selection against many linked deleterious mutations ("background selection"). Recent evidence has suggested this may be generally important in a wide range of populations (see Hahn (2008) for a recent review). In this paper, we explore how this type of selection alters the clonal (i.e. allelic) structure of a population. Our analysis leads to a generalization of the Ewens sampling formula to situations involving background selection.

Over the past few decades, numerous authors have studied allelic diversity in infinite-alleles frameworks that incorporate selection. Li (1977) and Watterson (1978) introduced models in which alleles may have a few different selective effects. (Li 1978) and others (Li 1979; Ewens and Li 1980; Griffiths 1983) analyzed the structure of allelic diversity in these models. More recent work has analyzed a very general model of selection introduced by Ethier and Kurtz (1987), which allows for diverse types of selection pressures (Ethier and Kurtz 1994; Joyce and Tavare 1995; Grote and Speed 2002; Joyce 1995). This work has helped us understand the general effects of selection in distorting the frequency spectrum of sampled alleles.

However, the models these authors have analyzed cannot be directly connected to a concrete description of mutations and selection occurring at specific sites. Rather, they assume that each new mutation creates a new allele whose fitness is completely independent of the fitness of its parent. In other words, there is no sense of relatedness among alleles, or of a correlation in fitness between closely related alleles. Etheridge and Griffiths (2009) and Etheridge *et al.* (2010) have more recently derived a coalescent dual of the Moran process with an arbitrary number of types, mutation rates between types, and genic selection coefficients, but it is not clear how this corresponds to selection acting on some fraction of an infinite number of specific sites.

In this paper we take a different approach, based on the specific model of linked sites described by Charlesworth *et al.* (1993) and Hudson and Kaplan (1994). That is, we imagine that each individual has a genome comprised of many neutral and many negatively selected sites. The fitness of each individual is determined by the number of mutations it carries at the negatively selected sites. We make the infinite-sites assumption that no two mutations at the same site ever segregate simultaneously. This is also an infinite-alleles model, but it is based on a specific model of mutations at individual sites, and the fitness of each new allele depends on the fitness of its parent.

Earlier studies have investigated the effects of purifying selection in models identical or closely related to the one we consider here. Charlesworth *et al.* (1993) introduced a model essentially identical to the one we analyze here, and Kaplan *et al.* (1988) and Hudson and Kaplan (1994) developed a simple algorithm which can be used to recursively compute how purifying selection alters the structure of genealogies. Hudson and Kaplan (1995b) and Gordo *et al.* (2002) further developed this idea, resulting in a simple computational method for sampling genealogical relationships in the presence of background selection. Related simulation and analytical work has further characterized the structure of genealogies and the statistics of genetic diversity at the level of individual sites in this or closely related models (McVean and Charlesworth 2000; Seger *et al.* 2010; Charlesworth *et al.* 1993; Comeron and Kreitman 2002; Comeron *et al.* 2008; Barton and Etheridge 2004).

However, this earlier work does not provide an analytic description of lineage structure, or sampling formulae for allelic diversity in the presence of purifying selection on many linked sites.

In this paper, we explicitly analyze the lineage structure, and we derive a selected version of the Ewens sampling formula. We begin by noting that the balance between mutations at deleterious sites and selection against them leads to a steady state mutation-selection balance (Haigh 1978). Our approach is to study the structure of lineages within this steady state, using the Poisson Random Field (PRF) method developed by Sawyer and Hartl (1992). We show that this lineage structure can alternatively be derived using a retrospective approach, by considering the probabilities of mutation and coalescence events in the ancestry of each individual; these probabilities are calculated by Hudson and Kaplan (1994) and Gordo *et al.* (2002) (and implicitly in a related context by Barton and Etheridge (2004)). Our description of lineage structure is thus precisely consistent with the analysis of genealogical structures in this earlier work. Finally, we use our description of lineage structure to calculate sampling formulae for allelic diversity, and compare our predictions to the results of Monte Carlo simulations.

Provided that selection is strong and deleterious mutation rates are sufficiently small, our results show that the effect of background selection on allelic diversity is to reduce the effective population size without otherwise distorting the lineage structure. Our results are thus consistent with the effective population size approximation to background selection proposed by Charlesworth *et al.* (1993). For weaker selection, however, or higher mutation rates, the effective population size approximation breaks down, and the effects of background selection become more complex. We show that in this case the allelic diversity cannot be described by neutral theory with some appropriately chosen effective population size. This is consistent with earlier observations that background selection leads to distortions in the structure of genealogies (McVean and Charlesworth 2000; Seger *et al.* 2010; O'Fallon *et al.* 2010; Comeron and Kreitman 2002; Comeron *et al.* 2008; Barton and Etheridge 2004; Gordo *et al.* 2002; Hermisson *et al.* 2002; Williamson and Orive 2002). Our analysis here allows us to compute precisely how these distortions due to purify-

ing selection at many linked sites alter patterns of allelic diversity, and hence provides an analytical framework for exploring where statistical power may lie to distinguish purifying selection from neutrality.

Our approach relies on the assumption that we can describe the distribution of fitnesses within the population with the steady state mutation-selection balance. In particular, we neglect fluctuations within this balance. We note that the PRF and retrospective approaches depend somewhat differently on this key approximation, which offers some insight into the role of fluctuations in our model. We analyze the validity of this approximation in more detail below, and describe a correction for some aspects of the effects of fluctuations in the PRF formalism, which allows us to make a precise correspondence with the retrospective approach. Related to this approximation, we also neglect the effects of Muller's ratchet. We discuss this approximation in detail in the Discussion. We further test the validity of our analysis via Monte Carlo simulations; we find that these approximations are reasonable across a broad parameter regime spanning weak and strong selective pressures.

Our analysis in this paper is limited to allelic diversity, and it does not address the degree of relatedness among sampled alleles. In other words, our analysis only tells us the probability that individuals are genetically identical, not the distribution of the number of specific sites at which individuals may differ. Our results are thus not directly comparable to the work described above, which makes predictions about expected diversity at the level of individual sites. However, while our allele-based results provide an incomplete picture of genetic diversity within the population, they do provide a useful perspective on how purifying selection distorts patterns of molecular evolution. Most importantly, we are able to make precise analytical predictions about how purifying selection distorts allelic diversity, in ways that cannot be described by a single reduced effective population size.

3.2 Model

We imagine a finite haploid population of constant size N. Each haploid genome has a large number of sites, which begin in some ancestral state and mutate at a constant

rate. Each mutation is either neutral or confers some fitness disadvantage s (where by convention $s > 0$). We assume an infinite-sites framework, so there is negligible probability that two mutations segregate simultaneously at the same site.

We assume that there is no epistasis for fitness, and that each deleterious mutation carries fitness cost s, so that the fitness of an individual with k deleterious mutations is $w_k = (1-s)^k$. Since we assume that $s \ll 1$, we will often approximate w_k by $1-sk$. Later we comment briefly on extensions to our method to consider the case when the selection coefficient of a deleterious mutations is drawn from some fixed distribution.

The population dynamics are assumed to follow the diffusion limit of the standard Wright-Fisher model. That is, we assume that deleterious mutations occur at a genome-wide rate U_d per individual per generation (with deleterious mutations assumed to be decoupled from selection). We define $\theta_d/2 \equiv NU_d$, the per-genome scaled deleterious mutation rate. Similarly, neutral mutations occur at a rate U_n per individual per generation, and we analogously define $\theta_n/2 \equiv NU_n$. We assume that each newly arising mutation occurs at a site at which there are no other segregating polymorphisms in the population (the infinite-sites assumption). Since in this paper we focus only on allelic diversity, this infinite-sites approximation simply means that each new mutation creates a unique allele. Throughout the analysis we assume that Muller's ratchet can be neglected; we discuss the validity of this approximation in the Discussion.

We study the case of perfect linkage. In other words, we imagine that all the sites we are considering are in an asexual genome or within a short enough distance in a sexual genome that recombination can be entirely neglected. Although our model is defined for haploids, this assumption means that our analysis also applies to diploid populations provided that there is no dominance (i.e. being homozygous for the deleterious mutation carries twice the fitness cost as being heterozygous).

We believe that this is the simplest possible model based on a concrete picture of mutations at individual sites that can describe the effects of a large number of linked negatively selected sites on patterns of genetic variation. It is essentially equivalent to the model described by Charlesworth *et al.* (1993) and Hudson and Kaplan (1994), which has formed the basis for much of the analysis of background selection

(Gordo *et al.* 2002; Seger *et al.* 2010).

3.3 Analysis

The balance between mutations and selection leads to a steady state distribution of fitnesses within the population; this is the well-known 'mutation-selection balance'. However, the individuals of a given fitness are not all genetically homogeneous, but rather comprise a number of different alleles. The number and frequency distribution of these alleles depends on how quickly new alleles are created by deleterious mutations from more-fit individuals, and hence on the overall fitness distribution.

We begin by describing the relevant aspects of the mutation-selection balance that leads to a steady state distribution of fitnesses within the population. Our description of this steady state fitness distribution is entirely deterministic. Of course, in a finite population, there will be random fluctuations in the values of h_k, the fraction of the population harboring k deleterious mutations. In the most extreme case, these fluctuations lead to Muller's ratchet. In our analysis below, we will neglect these fluctuations in h_k, assuming that these frequencies are always at their deterministic steady state. Consistent with this approximation, we will also neglect the effects of Muller's ratchet. We will then return in a later section to use our results to determine when these approximations are valid.

If we assume for a moment that these approximations are reasonable, we can already guess the form of our result for the allelic diversity. New alleles are constantly being generated within fitness class k due to deleterious mutations from class $k-1$ and neutral mutations from class k. Within class k, all alleles drift *neutrally with respect to each other*. Therefore, conditional on mutations and selection keeping the frequency of the class at h_k, the allelic diversity *within this class* will be the same as in a neutral population of size Nh_k in which new alleles are created by mutations at the appropriate rate. Thus for example the probability two individuals are of the same allelic type is the probability that they are both in the same class k times the appropriate neutral result for the homozygosity within that class, summed over all possible classes. Sampling formulae for larger samples can be calculated in the

analogous way.

The remainder of our analysis in this paper is, essentially, devoted to making this simple intuition precise and showing when it is accurate. We start by summarizing earlier results for the steady state mutation-selection balance h_k, and then compute the allelic diversity in detail, neglecting all fluctuations in h_k. This allows us to see precisely when this approximation is reasonable, and hence prove when the simple intuition described above holds.

3.3.1 The Steady State Fitness Distribution

In our model, all deleterious mutations have the same fitness cost s, so we can characterize individuals by their Hamming class, k, relative to the wildtype (which by definition has $k = 0$). That is, individuals in class k have k deleterious mutations more than the most-fit individuals in the population. Here k refers only to the number of *deleterious* mutations an individual has; individuals with the same k can have different numbers of neutral mutations. We normalize fitness such that by definition all individuals in class $k = 0$ have fitness 1. Individuals in class k then have fitness $1 - ks$.

Imagine that at a given time a fraction $h_k(t)$ of the population is in class k. This class is acquiring new individuals due to deleterious mutations arising in class $k-1$, and it is losing individuals due to deleterious mutations away to class $k+1$. It also gains or loses individuals at a rate $-(k - \bar{k})s$ due to selection, where $\bar{k}$ is the mean k within the population, $\bar{k} \equiv \sum k h_k$. This is illustrated in Fig. 3.1. Note that the term involving $\bar{k}$ simply normalizes the effect of selection (selection favors a class if it is more fit than the average individual, and vice versa). This means that on average $h_k(t)$ will evolve according to the equation

$$\frac{dh_k(t)}{dt} = U_d h_{k-1} - U_d h_k - (k - \bar{k}) h_k s. \tag{3.1}$$

Note this is a system of k equations for all the $h_k(t)$. Of course random genetic drift will also affect the $h_k(t)$, and these deterministic equations are only true on average. We return to this point below, but for now we neglect drift and focus on the steady state distribution.

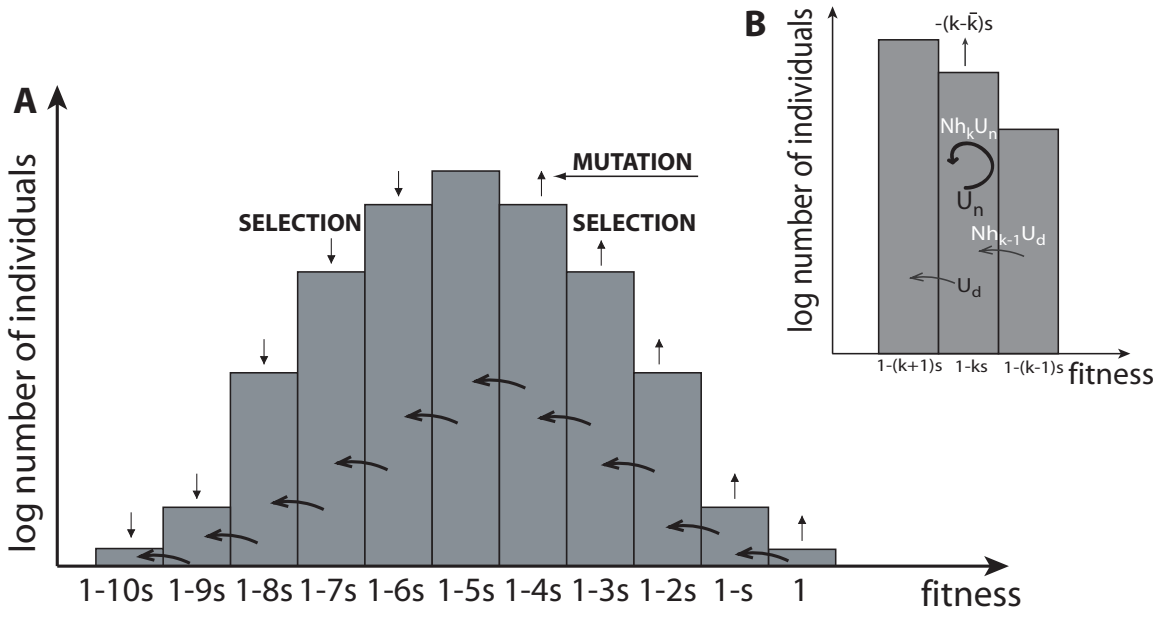

Figure 3.1: **Schematic of the Allelic Diversity in Mutation-Selection Balance: (a)** Sketch of the mutation-selection balance in the case $\frac{U_d}{s} = 5$. The steady state distribution of fitness within the population is maintained by a balance between mutations moving individuals towards lower fitness and selection favoring those classes more fit than average at the expense of those less fit than average. **(b)** The inset shows the processes maintaining a class of individuals with k deleterious mutations. Deleterious mutations from class $k-1$ found new lineages within class k at rate $Nh_{k-1}U_d$. Neutral mutations found new lineages in the class at a rate Nh_kU_n. Selection favors or disfavors individuals from each lineage at a per capita rate $-(k-\bar{k})s$, and deleterious mutations eliminate individuals from each lineage at a per capita rate $U_d + U_n$.

The steady state fitness distribution (the mutation-selection balance) is given by the values of $h_k(t)$ after a long time. We can find this mutation-selection balance by setting the right hand side of Eq. (3.1) equal to 0 for all values of k. This calculation was originally carried out by Kimura and Maruyama (1966) and Haigh (1978); they found that the steady state, $\hat{h}_k$, is given by a Poisson distribution with mean $\frac{U_d}{s}$,

$$\hat{h}_k = \frac{e^{-U_d/s}}{k!}\left(\frac{U_d}{s}\right)^k. \tag{3.2}$$

Note that this means the average fitness in the population is $1 - U_d$, and $\bar{k} = \frac{U_d}{s}$.

3.3.2 Allelic Diversity within a given Fitness Class

We now look more closely at individuals within a given fitness class, as illustrated in Fig. 3.1b. For the moment we neglect neutral mutations; we consider their effects further below.

All lineages in class k originally arose from a deleterious mutation to an individual in class $k-1$. Each of these deleterious mutations founds a new lineage within class k. Such lineages are founded at a rate $\theta_k/2$, where we define

$$\theta_k = 2Nh_{k-1}U_d. \tag{3.3}$$

Note this is true whether or not the h_k are at their steady-state values, though for the purposes of our analysis we will always assume the steady state.

In our infinite-alleles approximation, each new lineage is an allele that is unique within the population. The fate of this lineage (allele) is then determined by the forces of random drift, selection, and additional mutations. Additional mutations that occur within this lineage go on to found new alleles. Thus from the point of view of this particular lineage, additional mutations cause individuals to be lost from the lineage. This means that individuals are removed from a lineage in class k at a per capita rate

$$s_k \equiv -U_d - s(k - \bar{k}). \tag{3.4}$$

We refer to s_k as the *effective selection coefficient* against an allele in class k, because it is the rate at which any particular lineage in class k loses individuals (note we have defined signs such that $s_k < 0$). Note that s_k depends implicitly on the h_k through the term involving $\bar{k}$ (recall $\bar{k}$ is the average value of k, $\bar{k} \equiv \sum kh_k$). For convenience we will define the scaled effective selection coefficient γ_k by

$$\gamma_k = Ns_k. \tag{3.5}$$

Note that in steady state, when the fitness distribution h_k takes the mutation-selection balance form $\hat{h}_k$ derived above, $\bar{k} = U_d/s$ and the effective selection coefficient s_k is negative for all fitness classes with $k > 0$. This makes intuitive sense: each fitness class (except $k = 0$) is constantly receiving new individuals due to mutations.

Thus older individuals must on average die out, if the fitness class is to stay at a constant steady state size. The only exception is the $k = 0$ class, for which $s_k = 0$. This class drifts effectively neutrally, with its actual selective advantage relative to the mean exactly balanced by the loss of individuals due to deleterious mutations. For $k = 1$ we have $s_1 = -s$, and in general $s_k = -ks$. On the other hand, θ_k/h_k increases with k, reflecting the fact that the stronger selection against the larger-k classes is balanced by a larger influx of new deleterious mutations into these classes.

We can now incorporate the effect of neutral mutations. Each neutral mutation within an individual in class k creates a new lineage in class k. Thus we may simply redefine the rate at which new lineages are founded, giving

$$\theta_k \equiv 2Nh_{k-1}U_d + 2Nh_kU_n. \tag{3.6}$$

When the h_k's are in steady state this defintion simplifies to $\theta_k = 2Nh_k(sk + U_n)$. Each neutral mutation also causes an individual to be lost from the lineage it was in before the mutation, so we also redefine the effective selection coefficient

$$s_k \equiv -U_d - U_n + s(k - \bar{k}). \tag{3.7}$$

These neutral mutations are also reflected in Fig. 3.1b. Note that for all k, neutral mutations tend to increase θ_k, and make s_k more negative. In the presence of neutral mutations, even s_0 is negative.

We have seen that new lineages are founded within fitness class k at rate $\theta_k/2$, and then drift randomly subject to an effective selective pressure s_k. We now make the key assumption that each lineage is independent of all the others. This assumption is valid provided that no lineage ever becomes a substantial fraction of the overall population, which will be true whenever $N|s_k| \gg 1$ (i.e. all lineages are selected against strongly enough). A sufficient condition for this to hold in the bulk of the fitness distribution is simply $N(U_n + U_d) \gg 1$, and in fact our approximation will also hold even in some circumstances when this condition breaks down (we describe this further below).

3.3.3 Poisson Random Field Description of Lineage Structure

Using the independence assumption, we have reduced the problem of describing a lineage within a given fitness class to exactly the situation addressed by the Poisson Random Field model of Sawyer and Hartl (1992). Thus the frequency distribution of lineages (alleles) in fitness class k is a Poisson Random Field (PRF) with parameters θ_k and γ_k (where as before $\gamma_k \equiv Ns_k$). That is, the number of distinct lineages in class k segregating at a frequency between a and b in the entire population is Poisson distributed with mean

$$\int_a^b f_k(x)dx, \tag{3.8}$$

where

$$f_k(x) = \frac{\theta_k}{x(1-x)} \frac{1 - e^{-2\gamma_k(1-x)}}{1 - e^{-2\gamma_k}}. \tag{3.9}$$

This is equivalent to saying that the probability that there exists a lineage in class k with frequency between x and $x + dx$ is $f_k(x)dx$, for infinitesimal dx. Note that this PRF result implicitly assumes that θ_k and γ_k are constant (which requires constant h_k), and hence only describes the diversity in steady state.

This PRF description offers a convenient and well-established way to describe the lineage structure. It is similar in spirit to the diffusion result used by Ewens (1972) in his original computation of the neutral ESF. However, there is an important difference: Ewens' $f(x)$ was derived as the solution to the diffusion approximation to the K-allele Wright-Fisher process, in the limit of infinite alleles. This explicitly constrains all lineages to add to a total frequency of 1. The PRF does not impose this constraint. This makes it possible to compute a simple analytical expression for $f_k(x)$ in the presence of selection. However, it does involve an implicit approximation. In the Supplementary Appendix, we describe this approximation along with a way to relax it using an alternative branching process model to describe lineage structure.

3.3.4 The Self-Consistency Condition

It is clear from our PRF formulation above that the allelic diversity within each fitness class depends on the θ_k and γ_k, which in turn depend on the h_k. Yet the sum of the

frequencies of all the alleles within fitness class k *is*, by definition, h_k. In steady state, these two quantities must be equal. Verifying under what conditions these quantities are equal allows us to determine in what parameter regime the PRF formulation is self-consistent.

More specifically, we have derived the steady state value of h_k in Eq. (3.2),

$$h_k = \frac{e^{-U_d/s}}{k!}\left(\frac{U_d}{s}\right)^k.$$

When we plug these h_k into our PRF result, the summed allele frequencies according to the PRF must agree with steady-state value we used for h_k, for consistency. According to our PRF result, the sum of the frequencies of all the alleles in fitness class k is

$$h_k = \int_0^1 x f_k(x)\, dx. \tag{3.10}$$

Because Eq. (3.2) is equivalent to requiring $\theta_k/2 = |\gamma_k| h_k$ for all k (i.e. in steady state the net influx of individuals into a class must equal the average rate at which individuals within that class are lost), we can rewrite the self-consistency equation as

$$\frac{\theta_k}{2|\gamma_k|} = \int_0^1 x \cdot \frac{\theta_k}{x(1-x)} \frac{1-e^{-2\gamma_k(1-x)}}{1-e^{-2\gamma_k}}\, dx. \tag{3.11}$$

Some algebra reduces this to the condition

$$\int_0^1 \frac{1-e^{-2\gamma_k x}}{x}\, dx = \frac{1-e^{-2\gamma_k}}{2|\gamma_k|}. \tag{3.12}$$

The analysis in Appendix A shows that this condition holds to the level of approximation considered whenever $|\gamma_k| \gg 1$. When this is true, the steady state mutation-selection balance of Eq. (3.2) is also the distribution h_k that makes our PRF analysis of the allelic diversity within each fitness class self-consistent.

The condition $|\gamma_k| \gg 1$ corresponds to saying that the effective selection coefficient in each class is large compared to $1/N$. This will be true for all k whenever $NU_n \gg 1$. In practice, even when this condition fails in some fitness classes, it is still valid for all classes in which $|\gamma_k| \gg 1$. Thus our results still give a good approximation to the population allelic diversity provided $|\gamma_k| \gg 1$ for the classes around $\bar{k}$ that make up the bulk of the population. This will hold whenever $\gamma_{\bar{k}} = N(U_d + U_n) \gg 1$.

When this condition does not apply, our PRF result for the allelic diversity within each fitness class is inaccurate. This is because, when $|\gamma_k| \gg 1$, the growth of some mutant lineages is limited by the size of the population, which is ignored by the PRF approximation. Thus the PRF approximation overestimates the probability that lineages become common, and the self-consistency breaks down.

It is important to note that we also require an additional, stronger condition for other aspects of our analysis to be valid. The self-consistency condition ensures that the average size of the fitness class implied by the PRF analysis equals the steady state h_k. However, even when this holds, there could be substantial fluctuations in h_k around its average value. The PRF result for $f_k(x)$ tells us the probability that a set of lineages exists at any given frequencies. Therefore it contains detailed information about these fluctuations. However, we have neglected these fluctuations in substituting the h_k into our expressions for θ_k and s_k, and will also neglect these fluctuations below in calculating sampling formulae. We return to consider this additional approximation in a later section.

3.3.5 An Alternative, Retrospective Approach

It is possible to derive the neutral Ewens sampling formula in two quite different ways. Ewens (1972) imagined new alleles being created continuously by new mutations, and considered the frequency distribution of lineages set up by the balance between the continual creation of new alleles and the extinction of older alleles. This leads to expressions analogous to those in our PRF calculation of the lineage structure. We can calculate sampling formulas from this lineage structure, as Ewens did in the neutral case. First, however, we note that in a companion paper to Ewens (1972), Karlin and McGregor (1972) showed that the Ewens sampling formula could also be derived using a retrospective analysis, by considering the ancestral history of a sample of individuals. This same type of retrospective approach is also possible in our model; in this section we describe this alternative derivation of the allelic diversity as relevant to the case of purifying selection.

In order to calculate the probability of a particular allelic configuration, we con-

sider the ancestral history of a sampled set of individuals. In particular, we are interested in the most recent event to occur in the history of a sample, backwards in time. We classify these possible events into one of three possible types: coalescence events (i.e. identity by descent), neutral mutations, and deleterious mutations.

This method is easiest to understand if we begin by considering a sample of size two. In order for two individuals to have the same genotype, they of course must be in the same fitness class k. Furthermore, if we look at the ancestral history of each of these two individuals, the most recent event to occur, backwards in time, must be a coalescent event. In contrast, for them to have a different genotype, the most recent event to occur must be a mutation event. Therefore, to calculate the probability of either configuration, we need only calculate the probability that the most recent event is a coalescent event.

In order to calculate the probabilities of each possible most recent event, we must know the distribution of times until each type of event. In general, neutral mutations are exponentially distributed with rate U_n per generation. Assuming the steady state values for h_k, deleterious mutations are also exponentially distributed with rate sk per generation (Hudson and Kaplan 1994). Finally, within each class, coalescence occurs as a neutral process with rate $\binom{i}{2}$ per Nh_k generations. Therefore, for a sample of size 2, each of which are sampled from class k, we have that:

$$\begin{aligned} P(\text{1st Event: Coal.}) &= \int_0^\infty dt P(\text{Coal at t}) P(\text{No Neut. Mut by t}) P(\text{No Del. Mut. by t}) \\ &= \int_0^\infty dt e^{-t} e^{-2Nh_k U_n t} e^{-2Nh_k skt} \\ &= \frac{1}{1+2Nh_k(U_n+sk)} = \frac{1}{1+\theta_k}, \end{aligned} \tag{3.13}$$

where we have defined $\theta_k \equiv 2Nh_k(sk+U_n)$. Of course, this result agrees with the standard neutral result, replacing θ by θ_k (see below).

This same logic can be easily extended to larger sample sizes. For example, if we consider i individuals within the same class, the probability that the first event is a coalescence event is

$$\begin{aligned} P(\text{1st Event: Coal.}) &= \int_0^\infty dt P(\text{Coal at t}) P(\text{No Neut. Mut by t}) P(\text{No Del. Mut. by t}) \\ &= \int_0^\infty dt \tbinom{i}{2} e^{-\tbinom{i}{2} t} e^{-iNh_k U_n t} e^{-iNh_k skt} \\ &= \frac{\binom{i}{2}}{\binom{i}{2} + iNh_k(U_n + sk)} = \frac{i-1}{i-1+\theta_k}. \end{aligned} \tag{3.14}$$

If the first event is a coalescence event, that means two of the individuals are of the same allelic type. This leaves us with $i-1$ individuals in the class which may or may not be identical; we can now use the identical method to ask whether any of these remaining individuals are of the same allelic type. Similarly, if the first event is a mutation event, the remaining $i-1$ individuals could still coalesce with each other before they also experience mutation events.

We note that our analysis in this section is very similar in spirit to that of Hudson and Kaplan (1994), Barton and Etheridge (2004), and particularly to Gordo *et al.* (2002). These earlier authors considered the relative probabilities of mutations and coalescence in the ancestry of each individual, leading to expressions that implicitly contain results analogous to those in this section. They did not however consider the implications of these results for the overall patterns of allelic diversity in the population, which we now turn to.

3.3.6 Sampling Formulae

We can now calculate the probability of sampled configurations of allelic types. Our goal is to calculate the probability that a sample of n individuals will have some distribution of allelic types (e.g. n_1 individuals with allele 1, n_2 individuals with allele 2, etc.). Specifically, we aim to calculate a negative selection version of the neutral Ewens sampling formula (ESF). As we will see, this calculation proceeds exactly analogously whether we use the lineage structure (PRF) or retrospective analysis.

We begin with the simplest case, a sample of $n = 2$ individuals from the population. What is the chance that these individuals are the same genotype? In other words, what is the allelic homozygosity, Q_2, in the population? In order to be the same genotype, the two individuals must carry the same number of deleterious mutations — i.e. they must fall in the same Hamming class, k. In addition, they must

also be of the same mutant lineage within class k. This must equal the probability that the most recent event in the history of these 2 individuals is a coalescence event; from Eq. (3.14) this is $\frac{1}{1+\theta_k}$. Alternatively, we could calculate the probability the two individuals are in the same lineage directly from our PRF result; it is the expected value of x^2, where x is integrated over the distribution of lineage frequencies in class k:

$$\int_0^1 \frac{x^2}{h_k^2} f_k(x) dx = \frac{1}{1+\theta_k}, \tag{3.15}$$

where we have evaluate the integral as described in Appendix A (see also the corrections in the Supplementary Appendix).

We therefore find that the full probability that two sampled individuals have the same genotype, which we denote Q_2, is given by

$$Q_2 = \sum_{k=0}^{\infty} h_k^2 \left(\frac{1}{1+\theta_k} \right). \tag{3.16}$$

Note that, in the case $U_d = 0$, all individuals are in the zero class, such that $h_{k\neq 0} \to 0$ and $h_0 \to 1$. Therefore:

$$Q_2^{Neutral} \to \frac{1}{1+2NU_n}, \tag{3.17}$$

in agreement with the neutral Ewens sampling formula.

In order for two individuals to have a different genotype, there are two possibilities: either the two individuals could be sampled from different classes (in which case they must have a different genotype), or they could be sampled from the same class, and be of different allelic types (cf. the first event in their ancestral history is a mutation event). Therefore:

$$Q_{1,1} = \sum_{k,k' \neq k} h_k h_{k'} + \sum_k h_k^2 \left(\frac{\theta_k}{1+\theta_k} \right) = 1 - \sum_k h_k^2 \left(\frac{1}{1+\theta_k} \right) = 1 - Q_2. \tag{3.18}$$

Note that:

$$Q_{1,1}^{Neutral} \to \frac{2NU_n}{1+2NU_n}, \tag{3.19}$$

in agreement with the neutral Ewens sampling formula.

Relationship with the Neutral Result:

At this point, it is informative to consider the form of this result. The presence of selection serves to subdivide the population into classes, as given by the mutation-selection balance result. Thus, in order for a sample of individuals to have a particular allelic configuration, they must be sampled from a set of classes consistent with that configuration. However, within each class, the population behaves identically to that of a neutral population, with a different population size ($N \to Nh_k$) and mutation rate ($U_n \to U_n + sk$). We can see this explicitly by defining:

$$Q^{ESF}_{\{Configuration\},k} \equiv \text{ESF Result for \{Configuration\} with } \theta \to 2Nh_k(U_n + sk). \tag{3.20}$$

For example, we have that:

$$Q^{ESF}_{\{2\},k} = \frac{1}{1+\theta_k}, \qquad Q^{ESF}_{\{1,1\},k} = \frac{\theta_k}{1+\theta_k}. \tag{3.21}$$

We can then rewrite our results as:

$$Q_2 = \sum_k h_k^2 Q^{ESF}_{\{2\},k}, \tag{3.22}$$

$$Q_{1,1} = \sum_k h_k^2 Q^{ESF}_{\{1,1\},k} + \sum_{k,k' \neq k} h_k h_{k'}. \tag{3.23}$$

Thus we see that, within each class, the probability of a particular configuration is effectively neutral with parameter $\theta = 2Nh_k(U_n + sk)$, consistent with our initial intuitive guess for the form of our result. The overall probability of a given allelic configuration is then the probability that a specific configuration is achieved within each class, summed over all possible sets of class configurations that are consistent with the allelic configuration.

Sample Size $n = 3$

This logic can be extended to larger sample sizes. In order for three randomly-selected individuals to have the same genotype, all three individuals must be sampled from the same class and they must all be from the same lineage (i.e. both of the first two

events must be coalescence). This can be computed by considering the average of x^3 over the PRF, $\int_0^1 x^3 f_k(x)dx$, or by using the results from Eq. (3.14). We find:

$$Q_3 = \sum_k h_k^3 \left(\frac{2}{2+\theta_k}\right)\left(\frac{1}{1+\theta_k}\right). \tag{3.24}$$

Note that, for $U_d = 0$, $h_{k\neq 0} \to 0$ and $h_0 \to 1$, such that:

$$Q_3^{Neutral} \to \frac{2}{(2+\theta)(1+\theta)}, \tag{3.25}$$

in agreement with the neutral Ewens sampling formula.

In order for two individuals to have the same genotype and the third individual to have a different genotype – a configuration we term bizygotic – there are two possibilities. First, two individuals could have been selected from the same class and the third individual could have been selected from a different class. In this case, the two individuals in the same class must be from the same lineage (i.e. coalesce prior to a mutation event). Alternatively, all three individuals could have been selected from the same class. In this case, two must be from the same lineage and the third from a different lineage, which occurs with probability

$$\int_0^1 3x^2(1-x)f_k(x)dx. \tag{3.26}$$

Thinking about this retrospectively, this is equivalent to the sum of two possibilities: either the first event could be a mutation event, in which case the next event among the other two lineages must be a coalescent event, or the first event could be a coalescent event, in which case the next event among the third lineage and the merged lineage must be a mutation event. We find

$$\begin{aligned} Q_{2,1} &= \textstyle\sum_{k,k'\neq k} 3h_k^2 h_{k'}\left(\frac{1}{1+\theta_k}\right) + \sum_k h_k^3\left[\left(\frac{2}{2+\theta_k}\right)\left(\frac{\theta_k}{1+\theta_k}\right) + \left(\frac{\theta_k}{2+\theta_k}\right)\left(\frac{1}{1+\theta_k}\right)\right] \\ &= \textstyle\sum_k \frac{3h_k^2}{1+\theta_k}\left(1-\frac{2h_k}{2+\theta_k}\right). \end{aligned} \tag{3.27}$$

Note that:

$$Q_{2,1}^{Neutral} \to \frac{3\theta}{(1+\theta)(2+\theta)}, \tag{3.28}$$

in agreement with the neutral Ewens sampling formula for this configuration, which we call bizygotic.

Analogous considerations lead to the probability that all three individuals are of different allelic types,

$$\begin{aligned} Q_{1,1,1} &= \textstyle\sum_{k,k'\neq k,k''\neq k',k} h_k h_{k'} h_{k''} + \sum_{k,k'\neq k} 3h_k^2 h_{k'} \left(\frac{\theta_k}{1+\theta_k}\right) + \sum_k h_k^3 \left(\frac{\theta_k}{2+\theta_k}\right)\left(\frac{\theta_k}{1+\theta_k}\right) \\ &= 1 - \textstyle\sum_k 3h_k^2 \left(\frac{1}{1+\theta_k}\right) + \sum_k h_k^3 \left(\frac{4}{(1+\theta_k)(2+\theta_k)}\right) = 1 - Q_3 - Q_{2,1}, \end{aligned} \tag{3.29}$$

as expected. Note that

$$Q_{1,1,1}^{Neutral} = \frac{\theta^2}{(1+\theta)(2+\theta)}, \tag{3.30}$$

in agreement with the neutral Ewens sampling formula.

Relationship with the Neutral Result

As before, we define a class-specific version of the neutral Ewens sampling formula with $\theta \to 2Nh_k(U_n + sk)$:

$$Q_{\{Configuration\},k}^{ESF} \equiv \text{ESF Result for \{Configuration\} with } \theta \to 2Nh_k(U_n + sk). \tag{3.31}$$

In particular, we have that:

$$Q_{\{3\},k}^{ESF} = \frac{2}{(1+\theta_k)(2+\theta_k)}, \qquad Q_{\{2,1\},k}^{ESF} = \frac{3\theta_k}{(1+\theta_k)(2+\theta_k)}, \qquad Q_{\{1,1,1\},k}^{ESF} = \frac{\theta_k^2}{(1+\theta_k)(2+\theta_k)}.$$

Using these formulae, we can rewrite our results:

$$Q_3 = \sum_k h_k^3 Q_{\{3\},k}^{ESF}, \tag{3.32}$$

$$Q_{2,1} = \sum_k h_k^3 Q_{\{2,1\},k}^{ESF} + \sum_{k,k'\neq k} 3h_k^2 h_{k'} Q_{\{2\},k}^{ESF}, \tag{3.33}$$

$$Q_{1,1,1} = \sum_k h_k^3 Q_{\{1,1,1\},k}^{ESF} + \sum_{k,k'\neq k} 3h_k^2 h_{k'} Q_{\{1,1\},k}^{ESF} + \sum_{k,k'\neq k,k''\neq k',k} h_k h_{k'} h_{k''}. \tag{3.34}$$

Therefore, we again see that, within each class, the probabilities of a particular configuration are effectively neutral with parameter $\theta \to 2Nh_k(U_n + sk)$. The overall probability of a given allelic configuration is then the probability that a specific configuration is achieved within each class, summed over all possible class configurations that are consistent with the allelic configuration.

Sampling Formulae for Arbitrary Sample Size

We can extend this method to arbitrary sample size. For example, in order for a sample of n individuals to each have the same genotype, all individuals must be sampled from the same class. They must all be of the same allelic type, which occurs with probability $\int_0^1 x^n f_k(x)dx$. Or equivalently, the first event among the n lineages must be a coalescent event, the next event among the remaining $n-1$ lineages must also be a coalescent event, and so on. We find

$$\begin{aligned} Q_n &= \sum_k h_k^n \left(\frac{n-1}{n-1+\theta_k}\right)\left(\frac{n-2}{n-2+\theta_k}\right)\cdots\left(\frac{1}{1+\theta_k}\right) \\ &= \sum_k \frac{h_k^n}{\binom{\theta_k+n-1}{\theta_k}}. \end{aligned} \tag{3.35}$$

Note that:

$$Q_n^{Neutral} \to \frac{1}{\binom{\theta+n-1}{\theta}}, \tag{3.36}$$

in agreement with the neutral Ewens sampling formula.

In principle, this method can be extended to calculate the probability of any allelic configuration. Alternatively, we can use the relationship between these results and the neutral Ewens sampling formula to infer the probabilities. We found that, for the cases $n = 2$ and $n = 3$, we can write the probability of a given allelic configuration as the probability that, within each class, a particular configuration is achieved, summed over all sets of class configurations that are consistent with the allelic configuration. Similarly, we see that for Q_n:

$$Q_n = \sum_k h_k^n Q_{\{n\},k}^{ESF}, \tag{3.37}$$

where we have defined:

$$Q_{\{Configuration\},k}^{ESF} \equiv \text{ESF Result for \{Configuration\} with } \theta \to 2Nh_k(U_n + sk). \tag{3.38}$$

Using this logic, we can infer the probability of additional configurations. For example, in order to sample n individuals of one genotype and $n-m$ of another, there are two possibilities: First, m individuals could be sampled from class k and $n-m$

individuals could be sampled from another class k'. The probability of sampling in this manner is $h_k^m h_{k'}^{n-m}\binom{n}{m}$. Within class k, the probability of the m individuals having the same genotype is given by the neutral result $Q^{ESF}_{\{m\},k}$ with $\theta \to 2Nh_k(sk+U_n)$. Similarly, within class k', the probability of the $n-m$ individuals having the same genotype is $Q^{ESF}_{\{n-m\},k'}$. Alternatively, all n individuals could be sampled from the same class k. This occurs with probability h_k^n. The probability of m individuals having the same genotype and $n-m$ individuals having another is then given by $Q^{ESF}_{\{m,n-m\},k}$. Combining these results and summing over all sets of k and k', we have that:

$$Q_{m,n-m}=\sum_k h_k^n Q^{ESF}_{\{m,n-m\},k}+\sum_{k,k'\neq k} h_k^m h_{k'}^{n-m}\tbinom{n}{m} Q^{ESF}_{\{m\},k} Q^{ESF}_{\{n-m\},k'}. \tag{3.39}$$

Note, however, that if $m = n-m$ we must divide by two in the second term in the above expression, to avoid double-counting.

Extending this logic, we have that:

$$\begin{aligned} Q_{n-m-p,m,p}=&\sum_k h_k^n Q^{ESF}_{\{n-m-p,m,p\},k}+\sum_{k,k'\neq k} h_k^{n-m-p} h_{k'}^{m+p}\tbinom{n}{m+p} Q^{ESF}_{\{n-m-p\},k} Q^{ESF}_{\{m,p\},k'} \\ &+\sum_{k,k'\neq k} h_k^p h_{k'}^{n-p}\tbinom{n}{p} Q^{ESF}_{\{p\},k} Q^{ESF}_{\{n-m-p,m\},k'}+\sum_{k,k'\neq k} h_k^m h_{k'}^{n-m}\tbinom{n}{m} Q^{ESF}_{\{m\},k} Q^{ESF}_{\{n-m-p,p\},k'} \\ &+\sum_{k,k'\neq k,k''\neq k,k'} h_k^{n-m-p} h_{k'}^m h_{k''}^p \tbinom{n}{n-m-p,m,p} Q^{ESF}_{\{n-m-p\},k} Q^{ESF}_{\{m\},k'} Q^{ESF}_{\{p\},k''}. \end{aligned} \tag{3.40}$$

Note, however, that we must correct the above expression for overcounting if two or more classes require identical configurations (e.g. if $n-m-p=m=p$ we must divide the second through fourth terms in the above expression by 3 and the last term by 6). In general, the probability of any allelic configuration can be written as the sum over all possible class combinations that are consistent with a given allelic configuration, where the probability of each configuration within a class is given by the neutral result with $\theta \to 2Nh_k(sk+U_n)$. In the Supplement we provide a computer algorithm that performs this sum symbolically, for any allelic configuration $Q_{i,j,k,\ldots}$.

Note that, in the case $U_d = 0$, all individuals are sampled from the zero-class, such that $h_{k\neq 0} \to 0$ and $h_0 \to 1$. In this case, only the leading-order term will be non-zero in the above results. Therefore, the results reduce exactly to the neutral Ewens sampling formula.

3.3.7 Fluctuations in the Steady State h_k

Even when the self-consistency condition holds, the frequencies h_k will fluctuate about their steady state frequencies. However, both our PRF description of the lineage structure and our retrospective analysis assume that the fitness distribution is always in the steady state, h_k. We have previously studied this approximation in Walczak *et al.* (2012). Here we summarize our analysis of the validity of this approximation, as relevant for the present paper.

Each allele in class k can at most contain $\frac{1}{s_k}$ individuals; selection prevents any individual allele from becoming more common than this. The total number of individuals in the class is on average Nh_k. Thus when $Nh_k \gg \frac{1}{s_k}$, each fitness class contains many individual alleles. Thus we expect that the overall fluctuations in h_k should be negligible provided that this condition holds. This intuition can be made precise: we can calculate the variance in h_k in steady state from our PRF approach, or more easily from a branching process approximation described in the Supplementary Appendix. By computing $Var(h_k)/h_k$, we show that in fact the fluctuations in h_k are indeed negligible provided that

$$Nh_k s_k \gg 1. \tag{3.41}$$

In practice, this condition will often not hold in the high-fitness (and low-fitness) tails of the distribution. However, provided it holds in the center of the fitness distribution from which most individuals will be sampled (i.e. for those fitness classes near the mean), our approach will still give a good approximation to the population allelic diversity.

We note that in addition to assuming h_k are in their steady state values in defining θ_k and s_k for both the PRF and retrospective approaches, the PRF contains an additional implicit approximation. In writing the PRF sampling formulae, we assumed that, for example, the probability two individuals in class k come from a lineage of frequency x (given that lineage exists) is $\frac{x^2}{h_k}$. This assumes that h_k and x are independent quantities. That is, we assume that all the lineages in the class always add up to a frequency h_k (i.e., we neglect fluctuations in h_k). However, the existence of

a high-frequency lineage naturally implies that h_k is likely to be larger than average, and vice versa.

These correlations between the frequency of an individual lineage and the h_k do not pose a problem to our retrospective analysis, which never makes reference to lineages, but it does lead to small errors in the PRF results. We show in the Supplementary Appendix that these errors are negligible provided that fluctuations in h_k can be neglected (i.e. provided $Nh_k s_k \gg 1$). However, they do lead to small discrepancies between the PRF and retrospective results (and between the PRF results and the neutral ESF in the $U_d \to 0$ limit, since the neutral ESF is derived assuming a strict constraint on the total population size). Thus in the Supplementary Appendix we describe a method to correct for these effects, making the lineage-based and retrospective approaches to allelic diversity exactly equivalent. All of the above sampling formulae include this correction, as do all our figures.

As a result of fluctuations in the values of h_k, there will also be fluctuations in the value of the average class, $\bar{k}$. But these are negligible in the same situations that fluctuations in h_k are.

There is one additional extreme effect of fluctuations in h_k: a fluctuation in h_0 can lead to loss of this most-fit class, a process referred to as Muller's ratchet. We expect that, provided the ratchet does not click many times over the timescale in which individual lineages exist, this will not significantly affect the allelic diversity. Thus we have neglected the ratchet in our analysis. We return to consider this in more detail in the Discussion, and test the validity of our approximation with numerical simulations.

3.3.8 A Distribution of Fitness Effects of Deleterious Mutations

We have analyzed a model in which all deleterious mutations have the same fitness cost, s. However, in most real populations it is likely that deleterious mutations have a range of possible fitness effects. We could model this by assuming that the overall deleterious mutation rate is still U_d, but that deleterious mutations have a fitness cost between s and $s + ds$ with probability $\rho(s)ds$. That is, $\rho(s)$ is the distribution

of fitness effects of deleterious mutations.

In this more general situation, there is still a steady state distribution of fitness within the population. Generalizing our earlier notation, we can write this distribution as $h(k)$, where $Nh(k)$ is the steady state number of individuals with a fitness between sk and $(s+ds)k$, where s is the average fitness cost of a deleterious mutation and k is no longer constrained to be an integer. For certain $\rho(s)$ (e.g. an exponential distribution) it is possible to calculate $h(k)$ analytically, but even when this is not possible there does exist some steady state $h(k)$.

The basic ideas behind our analysis still apply in this more general situation. The rate at which new lineages within fitness "class" $h(k)$ are created is now

$$\theta(k)/2 = Nh(k)U_n + N\int_0^k h(k')\rho((k-k')/s)dk'. \tag{3.42}$$

The effective selection pressure against individuals in this class is

$$s(k) = U_n + U_d - (k-\bar{k})s. \tag{3.43}$$

Using these modified parameters, we can now apply our analysis as before; the distribution of lineage frequencies in class k is given by the PRF formula $f(k;x)$ with appropriate $\theta(k)$ and $s(k)$. We can then find sampling formulas as before — the only difference is that instead of summing over a discrete set of fitness classes, we must integrate over a continuous set of possible fitnesses. For example, we have $Q_2 = \int_0^\infty \int_0^1 x^2 f(k,x)dxdk$.

This extension of our model allows us to calculate the effects of more general forms of purifying selection on allelic diversity. However, there is a wide array of possible distributions $\rho(s)$, and using this more general form obscures the basic effects of selection. Thus in analyzing our results and comparing to simulations we focus on the simpler case in which all deleterious mutations have the same fitness cost s. This focus has the advantage of simplicity, and it allows us to explore more clearly how the strength of selection affects the patterns of allelic diversity.

3.3.9 Simulations

In order to check the validity of our analysis, we have performed simulations of a Wright-Fisher population. In our simulations, we consider a population of constant size N and keep track of the frequencies of all genotypes over successive, discrete generations. In each generation, N individuals are sampled with replacement from the preceding generation, according to the standard Wright-Fisher process (Ewens 2004) in which the chance of sampling an individual is determined by its fitness relative to the population mean fitness.

In each generation, a Poisson number of deleterious mutations are introduced, with mean NU_d, and a Poisson number of neutral mutations are introduced, with mean NU_n. The mutations are distributed randomly and independently among the individuals in the population (so that a single individual might receive multiple mutations in a given generation). Each new mutation is ascribed to a novel site, so that each mutation results in a new genotype.

Starting from a monomorphic population, all simulations were run for at least $\frac{1}{s}\ln(U_d/s)$ generations (or for at least several times N generations when $U_d/s < 1$), to ensure relaxation both to the steady-state mutation-selection equilibrium and to the PRF equilibrium of allelic frequencies within each fitness class. Appropriate relaxation to steady state was checked by extending the simulations and ensuring our results did not change. The final state of the population – i.e. the frequencies of all surviving genotypes – was recorded at the last generation, and Q_2 and $Q_{2,1}$ were calculated from these frequencies. This was repeated and averaged over 250 replicate simulations to produce the points shown in the figures.

Our simulations allowed for random fluctuations in the frequencies of each fitness class, as well as for Muller's ratchet. The ratchet did not proceed substantially for the simulations relevant for Fig. 3.3, except for the highest U_d point shown in that figure. However, it did proceed substantially in the simulations shown in Fig. 3.2, such that the most-fit individuals at the end of each simulation contained typically a few (for small U_d/s) to more than a dozen (for larger $U_d/s \sim 10$) deleterious mutations. We can see that, despite the effects of Muller's ratchet and fluctuations in the h_k, our

simulations are generally in excellent agreement with our theoretical predictions.

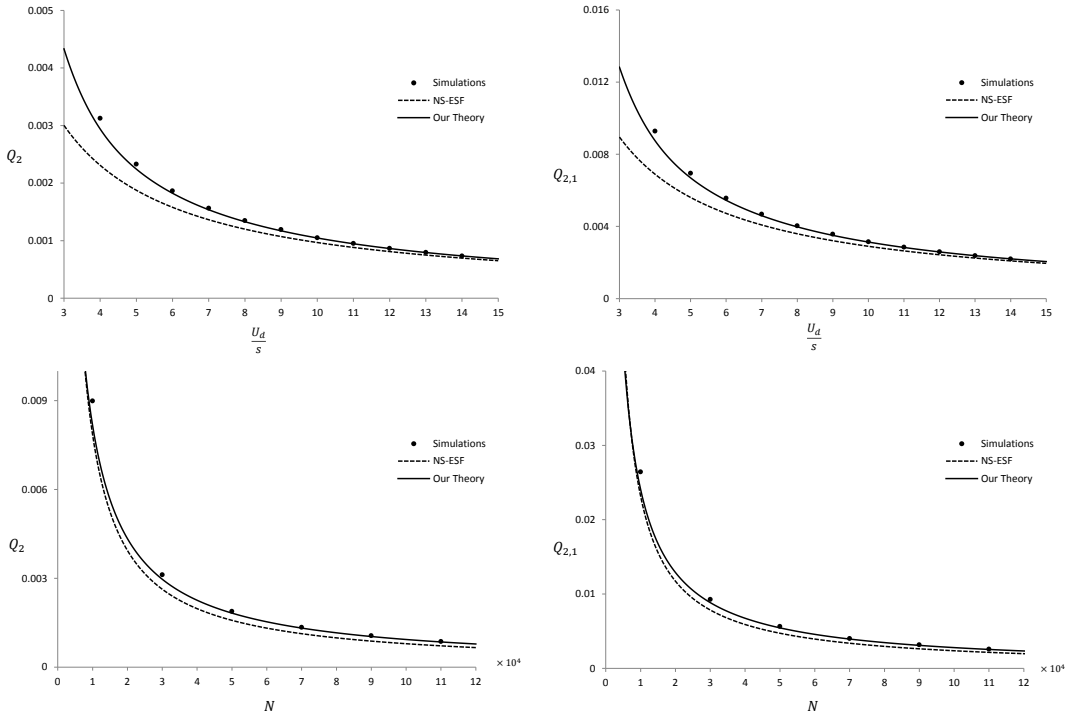

Figure 3.2: **A Comparison between Simulation Results (dots) and the Predictions of our Theory (gray lines),** for the case where some mutations are deleterious and others are neutral. For comparison we also show the predictions of NS interpretation of the neutral Ewens Sampling formula (black lines; the NM interpretation gives a worse fit to the data). **(a)** Homozygosity Q_2 as a function of U_d/s for $N = 5 \times 10^4$. **(b)** $Q_{2,1}$ as a function of U_d/s for $N = 5 \times 10^4$. **(c)** Homozygosity Q_2 as a function of N for $U_d/s = 6$. **(d)** $Q_{2,1}$ as a function of N for $U_d/s = 6$. In all plots $U_n = 3.2 \times 10^{-4}$, $s = 10^{-3}$.

3.4 Results and Discussion

Using the approach we have described, we can calculate the probability of any allelic configuration within a sample of n individuals from a population experiencing negative selection at many linked sites. From this, we can calculate the expected distribution of any statistic describing allelic diversity. To do so we must first determine

which allelic configurations lead to what values of the statistic. The probability of each possible value of the statistic is then the sum of the probabilities of all allelic configurations leading to that value. This is identical to the calculation we would do in the neutral case — the only difference is that to calculate the probability of each allelic configuration, we use our sampling formula rather than the neutral Ewens sampling formula.

In practice, some statistics are easier to calculate than others. While we can easily calculate the distribution of statistics describing diversity in a small sample, and we could in principle calculate certain statistics in larger samples (e.g. the total number of alleles in a sample of size n, K_n), further work is needed to develop efficient methods of calculating arbitrary statistics in large samples. This is clearly important for applications of our method to analysis of sequence data, but the combinatoric and computational issues involved are an extensive topic which is tangential to the ideas underlying our method. Instead, we focus here on describing the distributions of simple statistics involving small samples. Our aim is to highlight the essential differences between neutral diversity and the diversity in situations involving linked deleterious mutations.

Aside from likelihoods of configurations, and associated statistics, our approach could also be used to calculate the full distribution of branch lengths, following the generating function approach used by Lohse *et al.* (2011).

3.4.1 Relationship to the Neutral Ewens Sampling Formula

Although it may seem counterintuitive, our analysis applies even when $U_d = 0$ (that is, in the case where all mutations are neutral). In this case, our model is the same as that studied by Ewens (1972). If we apply our methods to this $U_d = 0$ case, all genotypes are in the fitness class $k = 0$, and we have $h_0 = 1$, $\gamma_0 = -NU_n$ and $\theta_0 = \theta = 2NU_n$. Provided that $|\gamma_0| \gg 1$, the conditions for our PRF analysis to be valid are met, and all of our previous results still apply, but are greatly simplified. And from our analysis of sampling formulas above we can immediately see that, as expected, setting $U_d = 0$ always causes our results to exactly reduce to the neutral

Ewens sampling formula. Note that we must take the limit $U_d \to 0$ rather than $s \to 0$ to recover the neutral result, because taking $s \to 0$ with finite U_d causes the steady state mutation-selection balance to break down (i.e. we have $h_k \to 0$ and fluctuations in the frequencies of each class become crucial).

For nonzero U_d, we expect that our results will differ from the predictions of the neutral ESF. To illustrate these differences in more detail, we study the allelic configurations in samples of size $n = 2$ and $n = 3$. Consider first the homozygosity Q_2 in a sample of size $n = 2$. In Fig. 3.2a and c we show how Q_2 depends on U_d and the population size N, both under our theory and in monte carlo simulations. We compare these results with the predictions of the neutral ESF. We make the same comparisons for the heterozygosity $Q_{2,1}$ in Fig. 3.2b and d. We note that the simulation results agree well with our predictions and differ from those of the ESF.

In making this comparison, there is some ambiguity about how to interpret the ESF, which depends only on θ, for $U_d > 0$. In one interpretation, we neglect selection against the deleterious mutations and set $\theta = 2N(U_n + U_d)$; we refer to this as the NS-ESF case. Alternatively, we could neglect the deleterious mutations entirely and set $\theta = 2NU_n$; we refer to this as the NM-ESF case.

In Fig. 3.3 we explore the ambiguity in the interpretation of the ESF, and compare the predictions of our theory to the two different interpretations of the ESF. For small U_d, our prediction is equivalent to both interpretations of the neutral ESF. As U_d increases, our predicted homozygosity decreases slowly until it experiences a sharp transition at $U_d \approx s$. This transition makes intuitive sense: when $U_d < s$, most individuals in the population have no deleterious mutations, and hence the allelic diversity is similar to the neutral case. As U_d increases past s, most individuals have deleterious mutations, so these mutations decrease the expected homozygosity. These deleterious mutations decrease homozygosity by less than they would if they were neutral, so our predicted homozygosity is higher than the NS-ESF (neglecting selection against deleterious mutations) but lower than the NM-ESF (neglecting deleterious mutations entirely).

We can gain further insight into this behavior by comparing our predictions to those of the NS-ESF and the NM-ESF in more detail (Fig. 3.3). We see that even

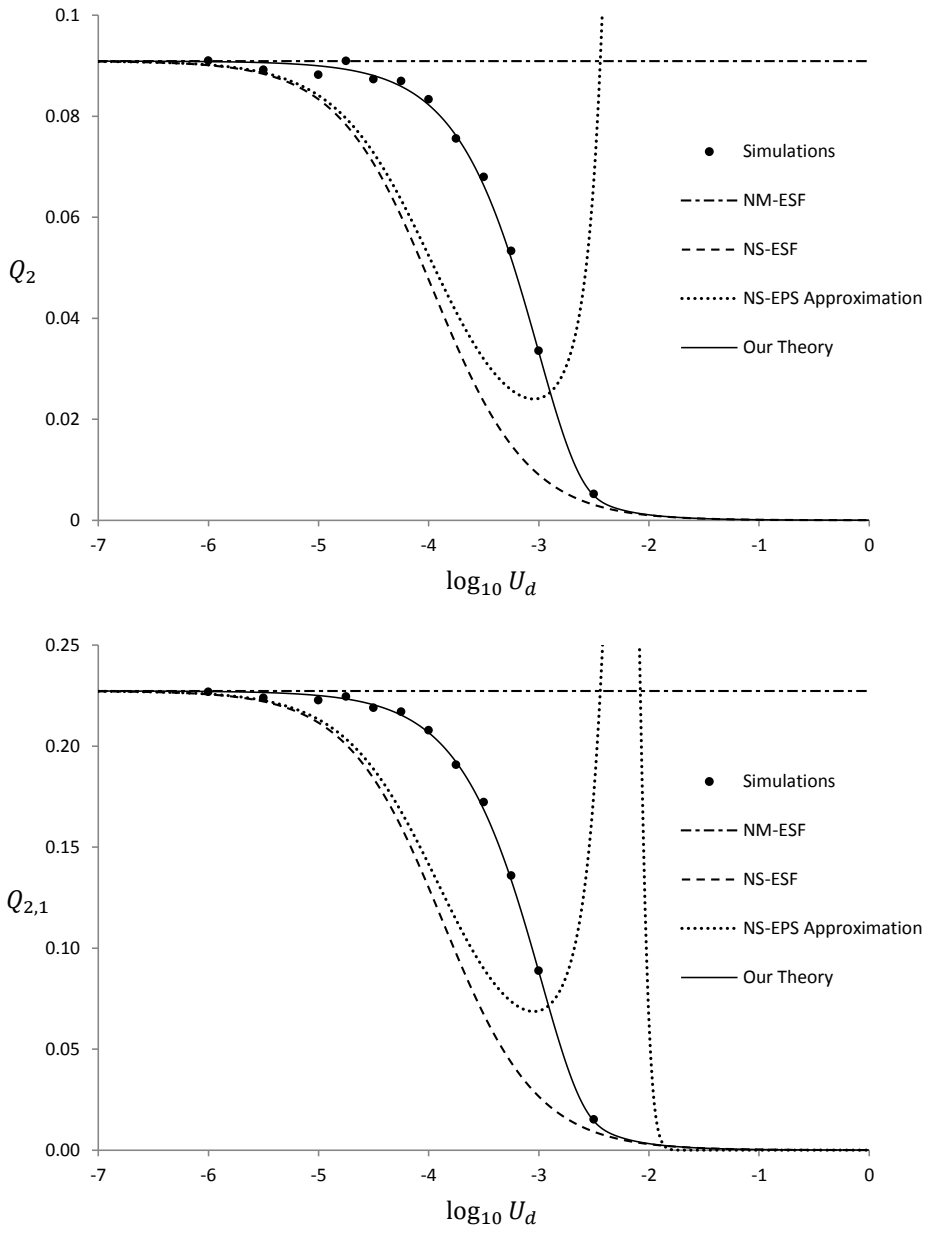

Figure 3.3: **Allelic Diversity as a Function of** $\ln U_d$**,** for $U_n = 10^{-4}$, $s = 10^{-3}$, and $N = 5 \times 10^4$. Our predictions are shown as a solid line, compared to the predictions of the NS-ESF (dotted line) and NM-ESF (dash-dotted line). We also compare our results to the predictions of a neutral ESF using the effective population size that would be predicted by background selection (BGS, dashed line), though we emphasize this is not the situation the BGS approximation was developed to address. These analytical predictions can be compared to simulation results (dots). **(a)** Homozygosity Q_2. **(b)** $Q_{2,1}$. Note that $Q_3 \approx 0$ everywhere for these parameters, so for these predictions $Q_{1,1,1} \approx 1 - Q_{2,1}$.

when $U_d = U_n$, our predicted homozygosity is only slightly lower than when $U_d = 0$, despite the fact that there are twice as many mutations occurring (and hence the NS-ESF prediction for Q_2 has declined by a factor of two). Here the NM-ESF prediction is fairly accurate, reflecting the fact that selection is still strong (with $U_d \ll s$) so that most individuals have no deleterious mutations at all. However, as U_d increases past s, most individuals now have one or more deleterious mutations and hence these mutations decrease our prediction for the allelic homozygosity. In this regime, the NM-ESF becomes inaccurate, because the deleterious mutations are sufficiently weakly selected ($U_d \gtrsim s$) that their presence is important to the diversity. However, despite this being weak selection, the fact that selection eliminates deleterious mutations from the population more rapidly than if they were neutral means that the allelic homozygosity is higher than the NS-ESF, even as U_d becomes very large. As U_d increases, our predictions become more similar to the NS-ESF, and in the limit of infinite U_d will equal the NS-ESF. In Fig. 3.3b we show the bizygosity $Q_{2,1}$ as a function of U_d. Through this parameter range Q_3 is small, and so $Q_{1,1,1} \approx 1 - Q_{2,1}$. As Fig. 3.3b shows, the dependence of bizygosity on U_d is similar to the behavior of heterozygosity, for essentially the same reasons.

This shift in our results from being approximately equal to the NM-ESF for small U_d to the NS-ESF for large U_d has an intuitive explanation from the form of our results for θ_k. For $U_d \ll s$, h_0 is close to 1, since most individuals have no deleterious mutations. In this class, we have $\theta_0 = 2Nh_0s_0 \approx 2NU_n$, the same as the θ for the NM-ESF. Since diversity within each class is neutral with the appropriate θ, in this $U_d \ll s$ regime the diversity is approximately that predicted by the NM-ESF. On the other hand, in the limit of very large U_d, h_k becomes sharply peaked about $k = U_d/s$, so almost all individuals have approximately the same fitness, and individual deleterious mutations change fitness by a negligible amount. Thus the diversity is approximately that predicted by the NS-ESF. This behavior is exactly as reflected in Fig. 3.3, with the transition between the two regimes occurring at $U_d \sim s$, as this analysis would predict.

Our analysis above makes it clear that the difference between weak and strong selection for the purpose of allelic diversity is set by whether s is small or large com-

pared to U_d. We have potentially three regimes of selection strength. For $Ns < 1$, selection is ineffective relative to drift, and we always have nearly neutral diversity. For $Ns > 1$, we can have weak, moderate, or strong selection. When $s \ll U_d$, we have weak selection as described above; the NS-ESF is accurate. When $s \lesssim U_d$, we have a "moderate selection" regime where the diversity generated by the deleterious mutations themselves can be important, and hence the NM-ESF is inaccurate. However selection is not so weak that the NS-ESF is accurate either; the selection against the deleterious mutations does reduce the amount of diversity they contribute. In this regime, neither interpretation of the Ewens neutral sampling formula provides an accurate prediction for allelic diversity. Finally, for $s \gg U_d$, we have a "strong selection" regime, where deleterious mutations are eliminated quickly from the population and hence do not contribute to diversity, and the NM-ESF is accurate. The NS-ESF is also accurate in this regime when $U_d \ll U_n$ but it will underestimate homozygosity when $U_d \gtrsim U_n$. Note that in Fig. 3.3 we show a case where $s > U_n$, so there is a regime where $s \gg U_d$ but $U_d \gtrsim U_n$ and hence the NM-ESF is accurate but the NS-ESF is not. Such a regime does not exist in the case $s < U_n$, but otherwise the same qualitative patterns exist for the same reasons.

3.4.2 Comparison to the Effective Population Size Approximation

The background selection model we have studied has been the subject of much earlier work, although this has largely been focused on the structure of genealogies in the presence of purifying selection, rather than allelic diversity (Hudson and Kaplan 1994, 1995b; Gordo *et al.* 2002; Seger *et al.* 2010). A particularly simple and useful approximation to the effects of background selection was developed by Charlesworth *et al.* (1993), Charlesworth (1994), and Charlesworth *et al.* (1995). This approximation is widely used to summarize the effects of background selection (Hartl 1988). We refer to it here as the effective population size approximation (EPS). The EPS analysis makes predictions about the the structure of genealogies and hence about genetic diversity at the level of individual sites, not just the allelic diversity we consider here. Further, it focuses on the genetic diversity

among neutral mutations only. Thus it is not directly comparable to our results in this paper. Despite this, we find it instructive to briefly examine how EPS compares to our results, if we apply it to predict allelic diversity. We stress that this is not the interpretation intended by Charlesworth *et al.* (1993) and does not provide a fair picture of its accuracy in general. Since EPS describes the structure of genealogies, we defer a detailed discussion of the accuracy of the EPS approximation and its relationship to our results to Walczak *et al.* (2012), where we calculate the structure of genealogies under our model.

The EPS approximation assumes that deleterious mutations are eliminated by selection quickly compared to the coalescence time between two individuals who do not have any such mutations. When this is true, almost all neutral mutations we observe occurred in individuals that did not have any deleterious mutations, because they have little time to occur in individuals that do have deleterious mutations before these individuals are eliminated by selection. Thus, according to the EPS approximation, the genetic diversity among neutral sites linked to negatively selected sites is exactly the same as the entirely neutral case, but with the population size N replaced by the size of the least-loaded (i.e. most-fit) class. That is, N is replaced by the effective population size

$$N_e = Nh_0 = Ne^{-U_d/s}. \tag{3.44}$$

Given this N_e, EPS predicts that any properties of neutral diversity are identical to those of coalescent theory with the appropriate N_e. Applying this to the allelic diversity, this predicts that the sampling properties of neutral alleles will be given by the classical Ewens' sampling formula, using $\theta = 2NU_nh_0 = 2NU_ne^{-U_d/|s|}$. Note this is effectively a NM-EPS case, which seems most natural. An alternative NS-EPS case can be defined using $\theta = 2N(U_n + U_d)h_0$; this leads to similar conclusions.

In the strong selection regime where $U_d \ll s$, most individuals are in the 0-class. Thus our analysis predicts that this class will dominate allelic diversity, which will be neutral with $\theta_0 = 2Nh_0s_0 = 2Ne^{-U_d/s}U_n$. Thus our analysis reduces exactly to the predictions of the NM-EPS in this regime. This is the regime in which the EPS approximation is expected to hold (Walczak *et al.* 2012), so our analysis reduces to

the EPS in the regime in which it should.

However, for the moderate and weak selection regimes, $U_d \gtrsim s$, the EPS prediction breaks down dramatically, consistent with the earlier observations of Nordborg *et al.* (1996) and Kaiser and Charlesworth (2009). We graph this prediction in Fig. 3.3 (using the NS interpration of the EPS, which provides a slightly better prediction than the NM interpretation). In this regime the EPS predicts that the neutral homozygosity increases dramatically, since the least-loaded class becomes negligible in size. However, the homozygosity is not so large in reality, as our predictions demonstrate. Rather, both neutral and deleterious variation among individuals that harbor one or more deleterious mutations is important. Our theory accounts for this effect, while EPS fails because the approximation that the coalescence time between individuals is dominated by the time in the least-loaded class breaks down.

We note that, contrary to the intuition one might be tempted to draw from EPS, having more deleterious mutations can never decrease allelic diversity. That is, if we fix all other parameters, simply having more deleterious mutations (i.e. increasing U_d) does not reduce heterozygosity. Certainly it reduces *neutral* heterozygosity, but accounting for all variation a population with a larger deleterious mutation rate will have more allelic heterozygosity.

3.4.3 Distortions in Allelic Diversity

The above discussion makes clear that for given population sizes, mutation rates, and selection strengths, purifying selection changes the probabilities of particular allelic configurations in a sample. However, this does not necessarily imply that selection leads to *distortions* in the patterns of genetic variation compared to the neutral case. In the neutral case, the probabilities of all allelic configurations in a sample are determined by a single parameter θ. This means that we can infer θ from a statistic which depends on the probabilities of one set of allelic configurations, and this θ then predicts the expected distribution of all other statistics describing genetic variation within the population, provided it is evolving neutrally.

Our discussion of the EPS approximation above makes clear that for sufficiently

strong selection, genetic diversity is not distorted relative to the neutral case. In this section, we show that for moderate to weaker selection (relative to mutation rates), there is no effective population size N_e which can describe genetic diversity in our model. As we noted in the Introduction, this is consistent with earlier observations that background selection leads to distortions in the structure of genealogies (McVean and Charlesworth 2000; Seger *et al.* 2010; O'Fallon *et al.* 2010; Comeron and Kreitman 2002; Comeron *et al.* 2008; Barton and Etheridge 2004; Gordo *et al.* 2002; Hermisson *et al.* 2002; Williamson and Orive 2002). Here we compute precisely how these distortions alter particular aspects of the patterns of allelic diversity. Our analysis in this section demonstrates one place in which statistical power exists to distinguish purifying selection from neutral processes at a reduced effective population size. Our framework can in principle be used to explore where such statistical power lies more generally, but we leave this more general question for future work.

In this section, we simply show that there is no effective neutral population size N_e to describe diversity in our model. To do this, it is sufficient to show that the effective θ that one would infer from one statistic predicts the incorrect values of other statistics. The simplest way to do this is to begin with the Q_2 we would predict given some set of parameters. We calculate the effective θ_e one would infer from this Q_2 using the neutral ESF (i.e. we choose θ_e such that $Q_2 = \frac{1}{1+\theta_e}$). We then calculate the neutral prediction for $Q_{2,1}$ (or Q_3) based on this θ_e. We compare this with our predictions for $Q_{2,1}$ (or Q_3) given the real parameters. The difference between these two predictions is a measure of the deviation from neutrality. We show this deviation from neutrality, expressed as the ratio of the neutral effective population size prediction to the actual result, for $Q_{2,1}$ in Fig. 3.4a and for Q_3 in Fig. 3.4b.

We see from Fig. 3.4 that negative selection distorts the allelic diversity away from high-frequency polymorphisms and towards lower-frequency polymorphisms, for a given level of overall heterozygosity. The effects are strongest when U_d is of order (or slightly larger than) s, and the distortion is stronger for smaller U_n and N.

These two simple statistics measuring deviations from neutrality demonstrate that there is no effective population size describing allelic diversity. These particular com-

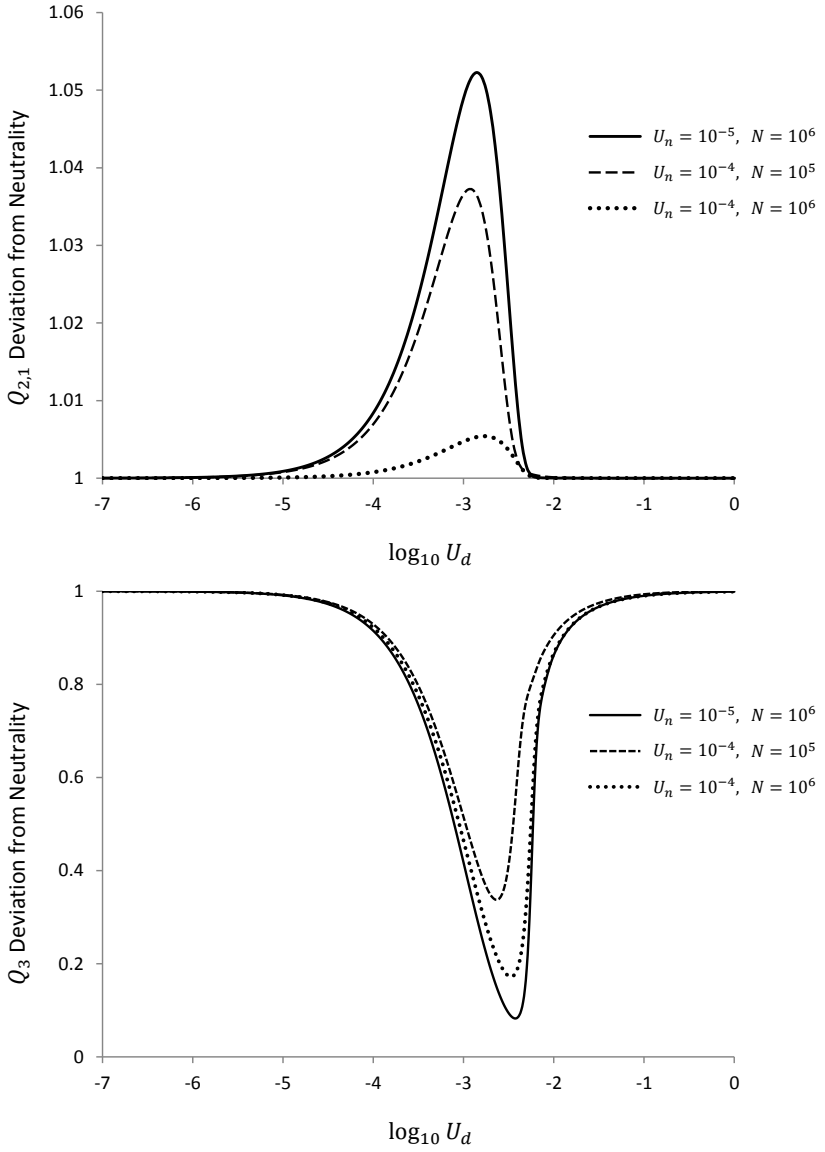

Figure 3.4: **The Deviation from Neutrality:** We take Q_2 as predicted by our theory, and use the neutral ESF to find the effective θ that this implies by setting $Q_2 = \frac{1}{1+\theta_e}$. We then use this effective θ_e in the neutral ESF to predict the values of $Q_{2,1}$ and Q_3 it corresponds to. We compare this to the $Q_{2,1}$ and Q_3 predicted by our theory. This is a measure of the deviation from neutrality, the skew in the frequency spectrum of allelic diversity away from neutral results with some modified effective population size. **(a)** The ratio of $Q_{2,1}$ from the effective population size description to the $Q_{2,1}$ from our theory, as a function of $\ln(U_d)$, for $s = 10^{-3}$ and three different values of U_n and N. **(b)** The ratio of Q_3 from the effective population size description to the Q_3 from our theory as a function of $\ln(U_d)$, for $s = 10^{-3}$ and three different values of U_n and N.

parisons are presumably not the most statistically powerful way to detect this type of negative selection, but they do show that statistical power exists. Using the framework developed in this paper, it is now possible to systematically investigate exactly how linked negatively selected sites generate different patterns of allelic diversity from the neutral case, and to determine which statistics provide the most power detect this type of selection. Note for example that the deviation from neutrality is much stronger in Fig. 3.4b than in Fig. 3.4a. This reflects the fact that we are inferring θ from Q_2, which in our theory is more closely related to $Q_{2,1}$ than it is to Q_3. Even more powerful tests for selection are presumably possible. While much earlier work has anticipated that purifying selection distorts the structure of genealogies (McVean and Charlesworth 2000; Gordo *et al.* 2002; Hahn 2008; Comeron *et al.* 2008; Seger *et al.* 2010; Betancourt *et al.* 2009; Comeron and Kreitman 2002; Williamson and Orive 2002), no analytic formalism has previously provided a way to determine precisely how selection alters patterns of allelic diversity (and hence, where statistical power may lie).

While we have shown that there is no neutral effective population size describing allelic diversity, this allelic diversity is a summary statistic of the full per-site diversity. Thus our result also implies that genetic diversity at a per-site level also cannot be described by a neutral effective population size, and that additional power to distinguish neutrality from negative selection can be found in data on site-based variation, consistent with the earlier work described above.

3.4.4 Muller's Ratchet

Throughout our analysis, we have assumed that Muller's ratchet can be neglected. This is clearly not true in general. The problem Muller's ratchet creates is that h_k can change with time, and this changes the distribution of allele frequencies within each class. After a "click" of the ratchet, the distribution of h_k shifts, eventually reaching a new state shifted left by one class (so the class that was originally at frequency h_k is now at frequency h_{k-1}, and so on). The PRF distribution of lineage frequencies in class k correspondingly shifts from f_k to f_{k-1}, and so on, which changes the allelic

diversity.

Fortunately, since $f_k(x)$ is similar to f_{k+1} and f_{k-1}, this effect is unlikely to cause major inaccuracies, provided the ratchet does not click many times over the timescale on which the lineage frequency spectrum turns over. We expect that this is generally true within the bulk of the fitness distribution. At the tails of the distribution, where h_k is small, the allele frequency distribution can sometimes be substantially different than expected due to the ratchet. However, by definition these classes represent a small fraction of the overall population and hence we do not expect them to contribute substantially to allelic diversity.

We tested the accuracy of our approximation neglecting Muller's ratchet using the simulations described above, all of which included the possibility of the ratchet. Our predictions remain very accurate, even in simulations in which the ratchet was observed to operate. Note, however, that the ratchet is potentially more problematic in considering the genetic diversity at the level of individual sites, because the high-fitness tail of the fitness distribution can be important for the structure of genealogies even if it does not contribute substantially to allelic diversity at any time.

3.4.5 Conclusion

We have introduced a formalism to calculate the statistics of allelic diversity in the presence of purifying selection at many linked selected sites. We have done so by calculating the structure of the individual lineages that maintain the deleterious mutation-selection balance. This analysis is based on the PRF framework of Sawyer and Hartl (1992), which was originally developed to describe the frequency of mutations at completely unlinked sites. We have adapted this framework to our problem with a shift in perspective: rather than treating new mutations at individual sites as the basic and independently fluctuating quantities, we consider the lineages founded by new mutations as the basic independent quantities. This allows us to describe aspects of the genetic diversity despite the fact that selection is acting on many linked non-independent sites. We showed that this approach is exactly equivalent to a retrospective perspective, which studied the probability individuals are in the same lineage

by considering the probability that coalescence events preceded mutations.

Of course, each lineage we describe contains many different mutations, and the fluctuations in lineage frequency described by the PRF framework represent correlated fluctuations in all of these individual mutations. If we could also describe how lineages are related to each other, and hence the statistics of which mutations they share, we could combine this with the results in this paper to describe the full per-site patterns of genetic diversity despite the correlations between sites introduced by linkage and selection. In this paper, however, we have focused on describing allelic diversity, leading to a negatively selected version of the neutral Ewens sampling formula. This analytical framework allows us to compute precisely how patterns of allelic diversity are distorted by negative selection at many linked sites, and hence understand exactly where statistical power may lie to distinguish purifying selection from neutrality.

3.5 Acknowledgements

We thank Warren Ewens and Isabel Gordo for helpful discussions and comments on the manuscript. MMD acknowledges support from the James S. McDonnell Foundation. LEN is supported by the Department of Defense through the National Defense Science and Engineering Graduate Fellowship Program, and also acknowledges support from an NSF graduate research fellowship. Many of the computations in this paper were run on the Odyssey cluster supported by the FAS Sciences Division Research Computing Group at Harvard University. AMW thanks the Princeton Center for Theoretical Science at Princeton University, where she was a fellow during some of her work on this paper. JBP acknowledges support from the James S. McDonnell Foundation, the Alfred P. Sloan Foundation, the David and Lucille Packard Foundation, the Burroughs Wellcome Fund, Defense Advanced Research Projects Agency (HR0011-05-1-0057), and the US National Institute of Allergy and Infectious Diseases (2U54AI057168).

3.6 Appendix A: Integrals involving $f_k(x)$

Our expressions for the probabilities of various allelic configurations involve integrals of the form

$$I = \int_0^1 A(x) f(x), \tag{3.45}$$

where $A(x)$ is a polynomial function of the form $A(x) = x^n(1-x)^m$ (with n and m integers). Here $f(x)$ is the expression from Eq. (3.9),

$$f(x) = \frac{ah}{e^a - 1} \frac{1}{x(1-x)} \left[e^{a(1-x)} - 1 \right], \tag{3.46}$$

where we have suppressed the subscripts and used the notation $a \equiv -2\gamma$.

Whenever n and m are both ≥ 1, these integrals are easy to evaluate analytically. When either n or m equals zero, the integrals can be separated into an exactly solvable analytical part and a part that involves the integral

$$I' = \int_0^1 \frac{e^{ay} - 1}{y} dy. \tag{3.47}$$

This integral I' is a known special function $Ein(-a)$; see p. 228 of Abramowitz and Stegun (1965).

Consider for example the integral

$$I_2 = \int_0^1 x^2 f(x) dx. \tag{3.48}$$

Substituting in for $f(x)$ and substituting $y = 1 - x$ in the integral gives

$$I_2 = \frac{ah}{e^a - 1} \int_0^1 \frac{1-y}{y} \left[e^{ay} - 1 \right] dy. \tag{3.49}$$

We now simply write $\frac{1-y}{y} = \frac{1}{y} - 1$ and evaluate the analytically solvable parts of this integral to get

$$I_2 = \frac{ah}{e^a - 1} I' - h + \frac{ah}{e^a - 1}. \tag{3.50}$$

Fortunately, we can calculate a simple analytic approximation for I' in the limit $a \gg 1$ (i.e. $|\gamma| \gg 1$), which is the limit we are always working in. This is a standard asymptotic expansion of the Ein function; we have

$$I' \approx \frac{1}{a} e^a \left[1 + \frac{1}{a} \right]. \tag{3.51}$$

We can now plug our approximation for I' into our result for I_2 to get

$$I_2 = \frac{h}{a}. \tag{3.52}$$

For more complex integrals, we need to keep higher order terms in the asymptotic expansion of I'. In general, we find

$$I_n = \int_0^1 x^n f(x) = \frac{(n-1)!h}{a^{n-1}}. \tag{3.53}$$

Similar calculations can be used to find an analogous approximation for $I_m = \int_0^1 (1-x)^m f(x)dx$, but this integral is not necessary for our purposes in this paper.

These calculations allow us to give simple analytic expressions for any integrals of the form $\int x^n(1-x)^m f(x)dx$. Whenever m and n are both ≥ 1, the integrals can be evaluated exactly in terms of elementary functions, and when either m or n are 0 we can use the above results to provide simple analytic approximations to whatever precision we require.

Chapter 4

Distortions in Genealogies due to Purifying Selection

Purifying selection can substantially alter patterns of molecular evolution. Its main effect is to reduce overall levels of genetic variation, leading to a reduced effective population size. However, it also distorts genealogies relative to neutral expectations. Hudson and Kaplan (1994) introduced a structured coalescent approach to describe this effect, which forms the basis for numerical methods and simulations. Here, we extend this approach by making the additional approximation that lineages may be treated independently, which is valid only in the strong selection regime. We show that in this regime, the distortions due to purifying selection can be described by a time-dependent effective population size and mutation rate, confirming earlier intuition. We calculate simple analytical expressions for these functions, $N_e(t)$ and $U_e(t)$. These results allow us to describe the structure of genealogies in a population under strong purifying selection as equivalent to a purely neutral population with varying population size and mutation rate, thereby enabling the use of neutral methods of inference and estimation for populations in the strong selection regime.

4.1 Introduction

Purifying selection purges deleterious mutations from a population, and hence reduces genetic variation at both selected and linked neutral sites. Charlesworth *et al.* (1993) introduced the background selection model to describe this effect. These authors observed that when selection is sufficiently strong, deleterious variants are quickly eliminated from the population, and thus all individuals are recently descended from individuals without deleterious mutations. Thus molecular variation is characteristic of a neutrally evolving population with a reduced effective population size. This simple and intuitive approximation — background selection reduces N_e — has been widely used to interpret patterns of molecular evolution in sequence data. We refer to it as the effective population size (EPS) approximation, and it successfully captures the dominant effect of strong purifying selection on the structure of genealogies: to decrease coalescence times without *distorting* genealogical structure.

However, even strong purifying selection does not act instantaneously. Instead, deleterious variants can segregate for a time that is inversely proportional to the strength of selection against them. This leads to two main distortions in the structure of genealogies. First, since purifying selection has not had time to act against deleterious mutations that occurred recently, the number of individuals that contribute to effective population size is higher in the recent past than the distant past. Numerous simulation and numerical studies have argued that this effect is similar to an effective population size $N_e(t)$ that declines as time recedes into the past (McVean and Charlesworth 2000; Comeron and Kreitman 2002; Gordo *et al.* 2002; O'Fallon *et al.* 2010; Seger *et al.* 2010; Walczak *et al.* 2012). Second, since individuals that acquired deleterious mutations in the distant past are less likely to have offspring in the present, mutations are not homogeneously distributed across genealogies. Recent work has argued that this effect can be summarized by an effective mutation rate $U_e(t)$ (representing the combined neutral and deleterious mutation rates) that also declines as time recedes into the past (Nielsen and Weinreich 1999; Woodhams 2006; O'Fallon 2010), though the potential importance of this effect is controversial (see Ho *et al.* (2011) for a recent review).

Recent evidence suggesting that purifying selection may substantially alter patterns of molecular evolution in nature (Eyre-Walker and Keightley 1999; Fay *et al.* 2001; Hahn 2008) has led to increased interest in understanding these effects. Several general theoretical approaches exist. The ancestral selection graph (Neuhauser and Krone 1997; Krone and Neuhauser 1997) offers a full formal solution, but is computationally unwieldy (Przeworski *et al.* 1999). An alternative approach is the structured coalescent method introduced by Kaplan *et al.* (1988). In this approach, the population is subdivided into classes of individuals at different fitnesses, where the average size of each fitness class is given by the steady state mutation-selection balance (Kimura and Maruyama 1966; Haigh 1978). In its most general form this method incorporates fluctuations in the class sizes, and hence can describe both weak and strong selection, but as a result is complex and requires numerical evaluation. This very general approach has since been further developed by Barton and Etheridge (2004) to address the effect of selection on genealogies at a linked neutral locus, including the effects of recombination. Hudson and Kaplan (1994) employed a simplified version of this structured coalescent method by approximating the distribution of fitness classes as fixed (i.e. neglecting fluctuations in their sizes). This leads to a simpler recursion describing the effects of purifying selection, which forms the basis for coalescent simulations (Gordo *et al.* 2002; Seger *et al.* 2010). We have recently shown that this recursion can be solved for the coalescence probabilities in each fitness class, leading to expressions for the structure of genealogies that can be evaluated numerically (Walczak *et al.* 2012). However, while these numerical and simulation methods offer important insight into the effects of selection on patterns of molecular evolution, they do not lead to simple analytic results.

In this paper, we propose an approximation which provides a simple analytic description of the leading effect of background selection in distorting genealogies. Our analysis provides an intuitive description of the main qualitative difference between a selected population and a neutral population with a reduced effective population size. Our results are necessarily more complex than the EPS result, since in addition to the main effect of background selection in reducing N_e they also capture the leading effect of background selection in distorting genealogies. However they are much sim-

pler (though correspondingly less generally valid) than the numerical and simulation methods of the full structured coalescent approach.

Our analysis is based on the simplified structured coalescent of Hudson and Kaplan (1994), which assumes the size of each fitness class is fixed at the steady state mutation-selection balance. We assume no recombination and neglect back mutations. We trace the ancestry of individuals as they move through the fitness distribution via mutations, as implemented in coalescent simulations by Gordo *et al.* (2002). We make the key additional approximation that the ancestry of each individual can be treated independently from all other individuals, which is valid only in the strong selection regime. We show that this implies that the structure of genealogies is equivalent to those in a neutrally evolving population with both a time-dependent effective population size and a time-dependent effective mutation rate, consistent with earlier intuition, and we calculate simple analytic formulas for $N_e(t)$ and $U_e(t)$. The time dependence in $N_e(t)$ reflects distortions in the structure of genealogies, while the $U_e(t)$ reflects the fact that mutations are not homogeneously distributed along the genealogies.

Our results are valid only within a limited parameter regime, and represent a special case of earlier more broadly applicable structured coalescent methods (Hudson and Kaplan 1994; Gordo *et al.* 2002; Barton and Etheridge 2004; O'Fallon *et al.* 2010; Seger *et al.* 2010). Our approximations highlight the conditions required for the effects of purifying selection to be summarized by an $N_e(t)$ and $U_e(t)$; when these conditions hold, the genealogies will be topologically neutral, and a selected population can be described as a neutral population with the appropriate time-varying population size and mutation rate. However, when these conditions fail, we expect selection to alter not only the distributions of coalescent branch lengths, but also the distribution of genealogical topologies.

We begin in the next section by reviewing the relevant aspects of the structured coalescent method of Hudson and Kaplan (1994), and discuss the approximations underlying this approach. We then calculate the ancestral fitness distribution, and use this to calculate the time-dependent effective population size $N_e(t)$ and mutation rate $U_e(t)$. We discuss the relationship between our results and the EPS approximation,

and compare our results with forward-time Wright-Fisher simulations. Finally, we describe how our results have potential practical applications in improving methods of inference and estimation for populations experiencing strong purifying selection. Most importantly, they make it possible to use preexisting neutral methods for inference of selection pressures, simply by using the appropriate $N_e(t)$ and $U_e(t)$. We also describe the implications of our results for understanding the potential role of purifying selection in explaining the apparently time-dependent mutation rates seen in recent experiments (Ho *et al.* 2005; Penny 2005; Burridge *et al.* 2008; Weir and Schluter 2008).

4.2 Model

We consider a haploid population of constant size N, with neutral mutation rate U_n and deleterious mutation rate U_d. Each deleterious mutation is assumed to confer a fixed fitness cost s, with $s \ll 1$. We assume no epistasis and multiplicative fitness, such that an individual carrying k deleterious mutations has fitness $(1-s)^k$. In this model, the population can be divided into fitness classes indexed by k. We assume an infinite-sites framework, such that all mutations introduce a new genotype into the population. We assume that there are no beneficial or back mutations, and we assume no recombination.

This model is equivalent to the mutation-selection balance framework described by Kimura and Maruyama (1966) and Haigh (1978). These authors showed that the fraction of the population in fitness class k, h_k, is given by

$$h_k = \frac{\left(\frac{U_d}{s}\right)^k e^{-U_d/s}}{k!}. \tag{4.1}$$

This is illustrated in Fig. 4.1.

We now summarize the structured coalescent method of Hudson and Kaplan (1994), as relevant for our analysis. Consider an individual sampled from fitness class k. Tracing the ancestry of this individual backwards in time, three types of events can occur. First, the individual may undergo a neutral mutation at rate U_n. Second, it can coalesce, but only with an individual in the same fitness class. Thus, it will

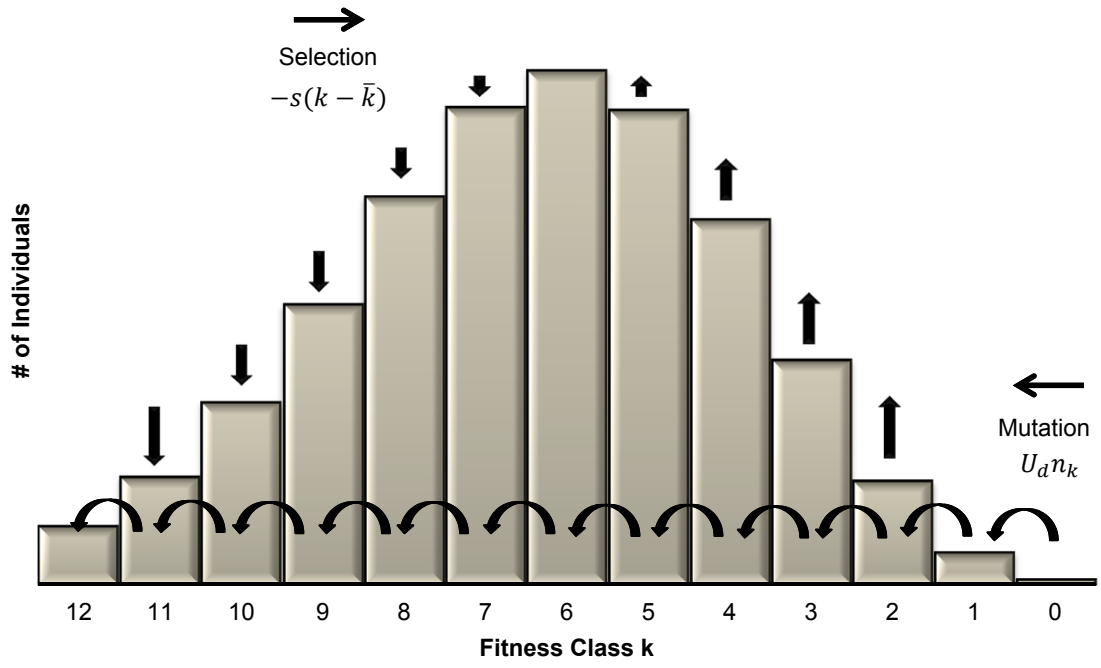

Figure 4.1: **: Schematic Depiction of Mutation-Selection Balance:** Deleterious mutations decrease the mean fitness of the population, while selection favors more-fit individuals. At steady state, a balance between these two effects is reached.

undergo coalescence with a specific individual from class k at rate $\frac{1}{Nh_k}$. Finally, it can undergo a deleterious mutation. Each generation, $Nh_{k-1}U_d$ individuals enter class k due to deleterious mutations from class $k-1$. Thus, the probability that an individual in class k underwent a deleterious mutation in the previous generation is approximately $\frac{Nh_{k-1}U_d}{Nh_k} = sk$. To summarize these possible types of events and their rates, we have

$$\text{Rates of Events} : \begin{cases} \text{Neutral Mutation} & U_n \\ \text{Deleterious Mutation} & sk \\ \text{Coalescence} & \frac{1}{Nh_k} \end{cases} \tag{4.2}$$

In this framework, each fitness class is treated as a subpopulation with size Nh_k and neutral mutation rate U_n. Within each class, all individuals are neutral with respect to one another. Deleterious mutation events are treated as migration events between the subpopulations. This migration occurs at rate sk, but may only occur in unit steps in one direction (towards higher fitness, backwards in time). This framework is equivalent to the diploid model used by Hudson and Kaplan (1994), for the case of no dominance.

This model makes use of an important approximation: we will assume throughout that the fraction of the population in fitness class k is fixed at the steady-state deterministic value, h_k. We refer to this as the steady state approximation. This approximation also implicitly neglects the effects of Muller's ratchet, which occurs when the zero-class fluctuates to extinction. In reality, the sizes of the classes will fluctuate due to random drift, and Muller's ratchet will occur. In general, the magnitude of genetic drift is inversely proportional to the population size. Thus, in order for the fluctuations in fitness class k to be negligible, we require that the magnitude of selection and mutation be large compared to the size of the class. This implies that our approximation will be reasonable provided $Nh_k sk \gg 1$. In a later section below, we show that our approximations are indeed valid in this parameter regime by comparing our results with forward-time Wright-Fisher simulations in which these fluctuations can occur and Muller's ratchet is able to proceed.

4.3 Analysis

4.3.1 The Ancestral Fitness Distribution

First, we calculate the ancestral fitness distribution for the population. We consider an individual sampled from class k. Deleterious mutations into the current class occur at a time exponentially distributed with rate sk. Then, at a time exponentially distributed with rate $s(k-1)$, the ancestral lineage will undergo the next deleterious mutation, and so on. In general, the probability that the ancestral lineage of an individual, sampled from class k_i in the present, mutated out of class k_f at time t in the past is the convolution of these steps

$$P_1(k_i \to k_f|t) = \int \delta(t - \sum t_j) \prod_{j=0}^{j=k_i-k_f-1} s(k_i - j)e^{-s(k_i-j)t_j} dt_j \tag{4.3}$$

The probability that the ancestral lineage remains in class k_f for time $t_{k_f-k_i}$ (i.e. does not undergo the next deleterious mutation) is $\int_{t_{k_f-k_i}}^{\infty} sk_f e^{-sk_f t'} dt' = e^{-sk_f t_{k_f-k_i}}$. By convolving these two results, we find in Appendix A the probability that an individual, sampled from class k_i in the present, was in class k_f at time t in the past,

$$P(k_i \to k_f|t) = e^{-sk_i t} \left(e^{st} - 1\right)^{k_i-k_f} \binom{k_i}{k_f}. \tag{4.4}$$

By summing over all possible starting classes k_i, weighted by their probabilities h_{k_i}, we find the probability that a randomly chosen individual was in class k_f at time t in the past,

$$\begin{aligned} p_{k_f}(t) &= \sum_{k_i=k_f}^{\infty} h_{k_i} P(k_i \to k_f|t) \\ &= \frac{\left(\frac{U}{s}e^{-st}\right)^{k_f} e^{-\frac{U}{s}e^{-st}}}{k_f!}. \end{aligned} \tag{4.5}$$

This is the ancestral fitness distribution of a randomly sampled individual; we illustrate it in Fig. 4.2. We note that, like the current fitness distribution, the ancestral fitness distribution is Poisson, but with reduced mean $\frac{U_d}{s}e^{-st}$. Thus, at time $t = 0$, the distribution is equivalent to the mutation-selection balance result. As $t \to \infty$,

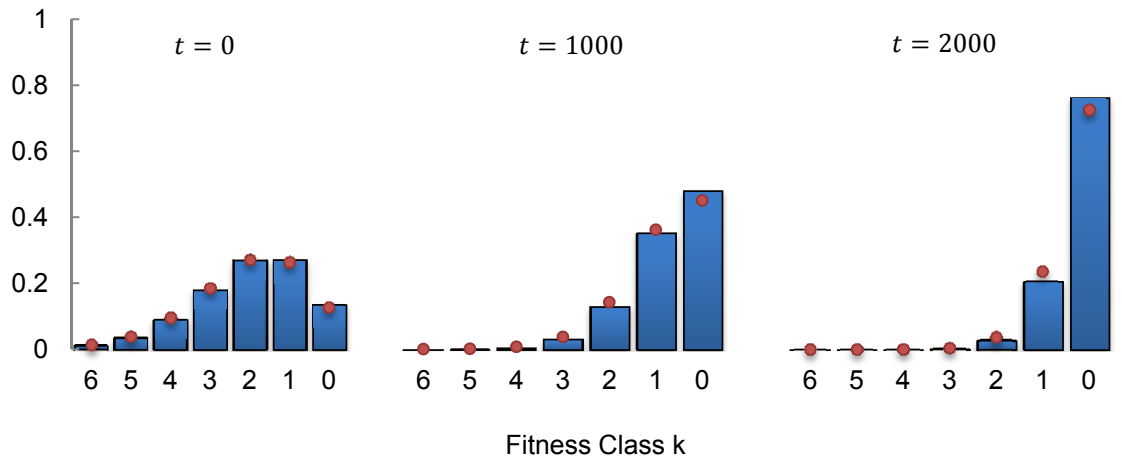

Figure 4.2: **The Ancestral Fitness Distribution,** shown for times $t = 0$, $t = 1000$, and $t = 2000$ before the present. The bars are the theoretical result, and the circles are simulations. In this plot, $\frac{U_d}{s} = 2$, $s = 0.001$, and $N = 10^5$. The present population is in mutation-selection balance. As time recedes into the past, the ancestral lineages shift towards higher fitness. As $t \to \infty$, all individuals eventually return to the zero-class.

the mean of the ancestral fitness distribution approaches zero, reflecting the fact that all individuals eventually descend from the zero-class.

This result intuitively agrees with the results of previous studies addressing the ancestral fitness distribution of a population under purifying selection (Hermisson *et al.* 2002; Barton and Etheridge 2004; O'Fallon *et al.* 2010). We find that the mean fitness of the ancestral lineages increases as time recedes into the past, $\langle k(t) \rangle = \frac{U_d}{s} e^{-st}$. Furthermore, the variance of the ancestral fitness distribution decreases as time recedes into the past, $Var[k(t)] = \frac{U_d}{s} e^{-st}$. The consequence of this is that ancestral individuals tend to have higher fitness, and tend to be in a narrower range of classes. This leads to significant consequences for both the apparent deleterious mutation rate and the per-generation probability of coalescence, as we will later see.

In the case of strong selection, the time to descend from the zero-class may be fast compared with a typical coalescence time within the zero-class (which is $Nh_0 =$

$Ne^{-U_d/s}$). This is the motivation behind the EPS approximation: if all individuals coalesce in the zero-class, and the time to descend from the zero-class is negligible in comparison to the coalescence time within the zero-class, then the population can be treated as a neutral population with size equal to that of the zero-class. However, as we will later see, the time to descend from the zero-class can be a significant fraction of the total coalescence time. This can lead to qualitative differences between a selected population and a neutral population with a fixed effective population size.

4.3.2 The Independent Lineage Approximation

The ancestral fitness distribution is defined for a single individual, moving through the distribution according to the probabilities described in Eq. (5.1). Our eventual goal will be to use this ancestral fitness distribution to understand the distributions of coalescence times among a sample of individuals. To do so, we will make the key approximation that the ancestral fitnesses of a larger sample of individuals can be drawn independently from this distribution. This approximation is analogous to a similar independence assumption made by O'Fallon *et al.* (2010).

In general, the ancestries of individuals will be correlated. In particular, by demanding that two or more lineages have not yet coalesced at a particular time, we bias the lineages to be further apart than average. Throughout our analysis, we neglect these biases. In general, if individuals are unlikely to share common ancestors except in the zero-class, and the time to coalesce is usually dominated by the time within the zero-class, then these distortions will not have a significant impact on the final result. Typical times to coalescence in the zero-class are of order $Ne^{-U_d/s}$, while deleterious mutation events through the distribution occur on a time-scale of $\frac{1}{sk}$. Thus, we can approximate lineages as independent provided $Nse^{-U_d/s} \gg 1$.

4.3.3 Effective Population Size

We now use the ancestral fitness distribution to compute per-generation coalescence probabilities. We have seen that two individuals in the same class will share a parent with probability $\frac{1}{Nh_k}$. Thus, as the ancestral fitness distribution shifts towards higher

fitness, ancestral individuals are more likely to be in the same class concurrently, and the per-generation probability of coalescence will increase over time. We can define a time-dependent effective population size as the inverse of the time-dependent per-generation coalescence probability,

$$P_n(t) = \frac{\binom{n}{2}}{N_e(t)}, \tag{4.6}$$

where $P_n(t)$ is the per-generation probability of coalescence in a sample of size n at time t. We show below that the $N_e(t)$ calculated in this manner is the same for any sample size within our framework.

Using the independence approximation, the probability that two ancestral individuals are each in class k at time t is $p_k(t)^2$. Therefore, we find for a sample of size two,

$$\frac{1}{N_e(t)} = \sum_{k=0}^{\infty} \frac{p_k(t)^2}{Nh_k} = \frac{e^{-\frac{2U}{s}e^{-st}}}{Ne^{-\frac{U}{s}}} \sum_{k=0}^{\infty} \frac{\left(\frac{U}{s}e^{-2st}\right)^k}{k!}.$$

This gives

$$N_e(t) = Ne^{-\frac{U}{s}(1-e^{-st})^2}. \tag{4.7}$$

Similarly, for arbitrary sample size, we have:

$$\begin{aligned} \frac{\binom{n}{2}}{N_e(t)} &= \sum_{k=0}^{\infty} \frac{1}{Nh_k} \left[\sum_{i=2}^{n} \binom{n}{i} \binom{i}{2} p_k^i \left(\sum_{k' \neq k} p_{k'} \right)^{n-i} \right] & (4.8) \\ &= \sum_{k=0}^{\infty} \frac{\binom{n}{2}}{Nh_k} \left[\sum_{i=0}^{n-2} \binom{n-2}{i} p_k^{i+2} \left(1-p_k\right)^{n-i-2} \right]. & (4.9) \end{aligned}$$

Using the binomial expansion:

$$(a+b)^n = \sum_{i=0}^{n} a^i b^{n-i} \binom{n}{i}, \tag{4.10}$$

and identifying $a = p_k$ and $b = 1 - p_k$, this becomes:

$$N_e(t) = Ne^{-\frac{U}{s}(1-e^{-st})^2}. \tag{4.11}$$

Thus, we see that there is a simple $N_e(t)$ that describes any size sample. In Fig. 4.3, we illustrate our analytical prediction for $N_e(t)$ and compare it to simulation results.

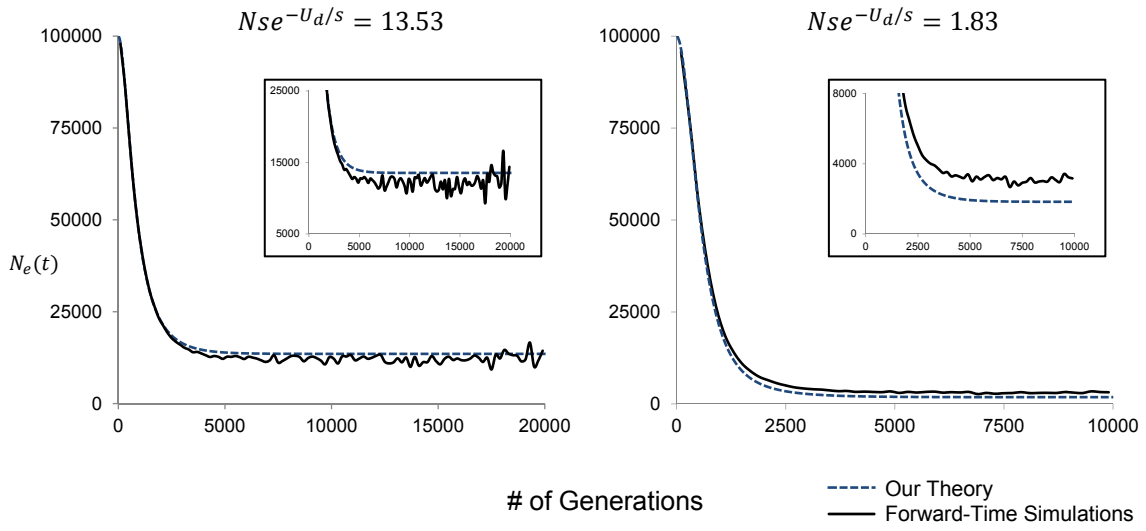

Figure 4.3: **Effective Population Size as a Function of Time,** for **(a)** $\frac{U_d}{s} = 2$ and **(b)** $\frac{U_d}{s} = 4$. In both cases, $N = 10^5$, $U_n = U_d$, and $s = 10^{-3}$. The effective population size begins at N, but undergoes a transition to a long-term rate of approximately $Ne^{-\frac{U_d}{s}}$.

We consider two parameter regimes. In the first, we have $Nse^{-U_d/s} = 13.53$, which represents a case where both the independent lineages and steady state approximations should hold reasonably well. In the second case, we have $Nse^{-U_d/s} = 1.83$, where both approximations begin to break down.

At $t = 0$, the effective population size is N. However, as $t \to \infty$, $N_e(t) \to Ne^{-U_d/s}$, reflecting the fact that all individuals will eventually return to the zero-class. At intermediate times, there is a transition between the initial and long term population sizes, representing the descent of lineages through the distribution. The rate of this transition depends primarily on the selection coefficient, s. We note that the EPS approximation corresponds to neglecting this transition, and assuming the long-time limit applies immediately.

The consequence of this time-dependent effective population size is that branch lengths in the recent past are relatively longer than branch lengths in the distant past. Thus, we are able to capture a distortion in the relative branch lengths within gene

genealogies. However, within our framework, the topologies of the genealogical trees are unchanged from neutral expectations. When the independence approximation does break down, it will break down more quickly for larger sample sizes, since the correlations among many individuals will be larger than among a pair of individuals. This means we no longer expect to find a single $N_e(t)$ for the whole population, and hence selection begins to distort tree topologies away from neutral expectations precisely at the point where our approximations break down.

4.3.4 Effective Mutation Rates

We now have a method for describing the structure of genealogies using a time-dependent effective population size. However, deleterious mutations will not be distributed homogeneously across these genealogies. We have seen that the rate of deleterious mutations, backwards in time, depends upon the current class of an individual. An individual at the mean fitness $\overline{k}$ has deleterious mutation rate $s\overline{k} = U_d$, as expected. An individual with more deleterious mutations is less fit, and thus will tend to die out from the population quicker. As a consequence of this, those individuals that do exist with a large number of deleterious mutations will have more recently descended from the previous class, such that their apparent mutation rate is higher. Analogously, individuals with fewer than average deleterious mutations will tend to die out more slowly, such that those who do exist appear to have a slower than average deleterious mutation rate. Thus, we see that the deleterious mutation rate depends upon the current class. A major consequence of this is that, as the ancestral fitness distribution shifts toward higher fitness, the effective mutation rate decreases. This captures the fact that deleterious mutations will be inhomogeneously distributed along genealogies, with a bias towards occurring more recently, as previously observed in simulations (Williamson and Orive 2002).

We can describe this effect by calculating the time-dependent rate at which mutations occur along the ancestry of a given individual. An individual in class k will undergo a deleterious mutation, backwards in time, at rate sk, and neutral mutations

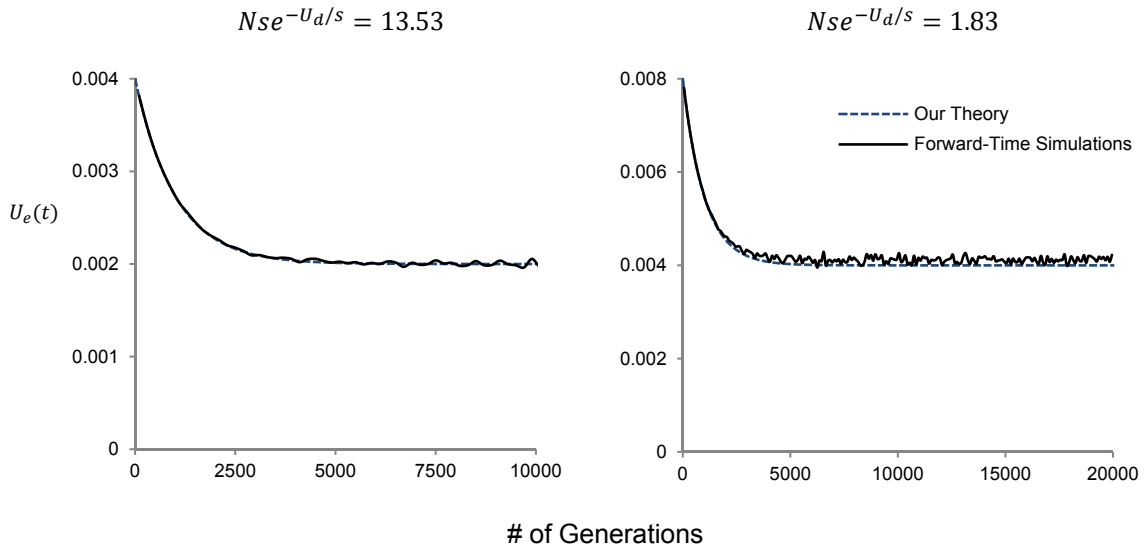

Figure 4.4: **Effective Mutation Rate as a Function of Time,** for **(a)** $U_d = 0.002$ and **(b)** $U_d = 0.004$. In both cases, $N = 10^5$, $U_n = U_d$, and $s = 10^{-3}$. The effective mutation rate begins at the instantaneous mutation rate, $U_n + U_d$, but undergoes a transition to a long-term rate of U_n. The transition is exponentially decreasing with rate given by the selection coefficient, s.

at rate U_n. Therefore we have

$$U_e(t) = \sum_{k=0}^{\infty} p_k(t)(U_n + sk) \tag{4.12}$$

$$= U_n + U_d e^{-st}. \tag{4.13}$$

In Fig. 4.4, we illustrate our prediction for $U_e(t)$ and compare it to simulation results, again using two different parameter regimes. At $t = 0$, the effective mutation rate is simply $U_n + U_d$, as expected. As $t \to \infty$, the mutation rate falls of to U_n, as in the EPS approximation. This is a consequence of the fact that for $t \to \infty$ all ancestral individuals have entered the zero-class, where only neutral mutations may occur backwards in time. More generally, this reflects the fact that if a deleterious mutation were to occur a long time in the past, it would be very likely to have died out, and thus not be sampled in the present. Therefore, the deleterious mutations that are seen in the present are biased toward more recent times.

4.4 Simulations

We performed forward-time Wright-Fisher simulations to confirm the validity of our results. Each generation, a new set of individuals were chosen from the previous set using multinomial sampling, and mutations were introduced as a Poisson process at rates NU_n and NU_d. The simulations ran for a total of at least $200,000$ $(2N)$ generations. These simulations allow for fluctuations in the class sizes, as well as for Muller's ratchet. In the parameter regime $N = 10^5$, $s = 10^{-3}$, $\frac{U_d}{s} = 4$, Muller's ratchet proceeded between 8 and 39 times in $200,000$ generations. In the parameter regime $N = 10^5$, $s = 10^{-3}$, $\frac{U_d}{s} = 2$, Muller's ratchet proceeded between 0 and 1 times in $200,000$ generations. The simulations were repeated at least 6000 times, and the results were averaged over trials.

4.5 Results and Discussion

Our analysis implies that the structure of genealogies in the presence of purifying selection is equivalent to a neutral population with the time-dependent effective population size $N_e(t)$ calculated above. Furthermore, we are able to account for the inhomogeneous distribution of mutations across these genealogies with our time-dependent effective mutation rate $U_e(t)$.

The idea that purifying selection can be described by a time-dependent effective population size is not new. For example, O'Fallon *et al.* (2010) also derived a time-dependent per- generation coalescence probability in the case of weak selection. They were able to calculate an ancestral fitness distribution using a continuous approximation, which is in turn used to calculate coalescent times. Other work by Seger *et al.* (2010) calculated a time-dependent effective population size by building upon the simulated structured coalescent approach of Hudson and Kaplan (1994) and Gordo *et al.* (2002). Our results are also based upon the framework of Hudson and Kaplan (1994), and should therefore be analogous to those above. Our work here is also related to our earlier analysis of the same model, in which we derived the distribution of simple statistics without making the additional independent lineages

approximation (Walczak *et al.* 2012; Desai *et al.* 2012). The analysis in this earlier work is more general, but does not lead to the simple analytical conclusions we reach here. We also note that Barton and Etheridge (2004) built on the more general structured coalescence approach of Kaplan *et al.* (1988) to calculate genealogical structure without assuming fitness classes are fixed in size, in a model where selection acts only on a single locus. Finally, we note that the concept of the ancestral fitness distribution was considered in detail in Hermisson *et al.* (2002). These authors derived the ancestral distribution for a set of haploid mutation-selection models, and our result can be seen as a limiting case of these results.

Although several of these earlier analyses found a time-dependent coalescence probability, none of them lead to simple analytical results describing precisely what $N_e(t)$ is. Although our analysis only holds in the strong-selection regime, we are able to account for qualitative differences between a selected population and the EPS approximation of a neutral population with reduced but constant effective population size $N_e = Ne^{-U_d/s}$, while maintaining an analytically simple formulation. Most importantly, we see that the $N_e(t)$ derived in this manner is the same for any sample size. Our result for $N_e(t)$ can therefore be used to calculate coalescence times among any sample from the population, provided the assumption of independent lineages can be maintained. Specifically, the distribution of the time to coalescence among a sample of size n is

$$\Psi_n(t) = \frac{\binom{n}{2}}{N_e(t)} e^{-\binom{n}{2}\int_0^t \frac{1}{N_e(t')}dt'}. \tag{4.14}$$

In Fig. 4.5, we compare this result with simulations, both in a parameter regime where our approximations are expected to hold and where our approximations are expected to break down. We also show for comparison the EPS approximation of Charlesworth *et al.* (1993), which assumes that all individuals are instantly descended from the zero-class. Our analysis is valid in a similar strong selection parameter regime, in which the time-scale of coalescence events is large compared to the time-scale of mutations through the distribution. However, we still account for the time of this descent, which leads to a qualitative difference between the predictions of the EPS approximation and our model — in particular, there is a non-zero peak in the

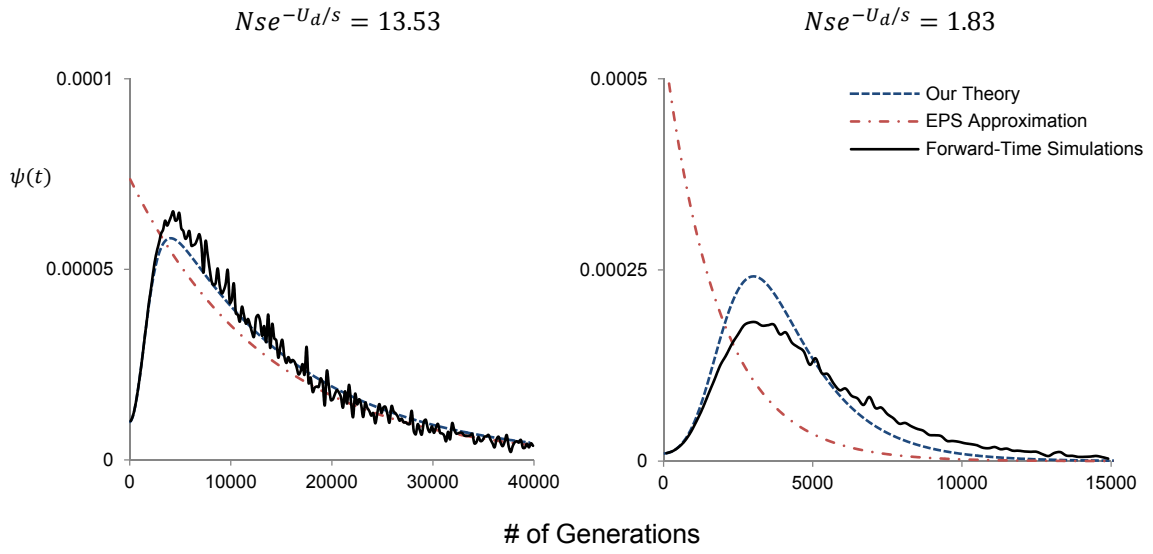

Figure 4.5: **Coalescence Probabilities as a Function of Time for a Sample of Size Two,** for **(a)** $\frac{U_d}{s} = 2$ and **(b)** $\frac{U_d}{s} = 4$. In both cases, $N = 10^5$, $U_n = U_d$, and $s = 10^{-3}$. In the effective population size (EPS) approximation, the per-generation coalescence probability is fixed at $\frac{1}{N_0}$, where $N_0 = Ne^{-\frac{U_d}{s}}$. Therefore, the probability of coalescence at a particular time is an exponentially decreasing function. In our theory, the per-generation coalescence probability begins at $\frac{1}{N}$, and then transitions to the long-time rate $\frac{1}{N_0}$. This introduces a non-zero peak in the overall probability of coalescence.

coalescence times reflecting the fact that the time to descend through the distribution makes coalescence at early times less likely. As $Nse^{-U_d/s} \to \infty$, our results approach the EPS approximation. For $Nse^{-U_d/s} \approx 1$ our approximation begins to break down, but it still partially captures the transition period in the coalescence probabilities and hence describes the qualitative features of the distribution of coalescence times more accurately than the EPS approximation.

We have shown that the distortions in genealogical structure due to a time-dependent effective population size are not the only qualitative effect of purifying selection on patterns of molecular evolution. We have also seen that deleterious mutations do not occur along these genealogies homogeneously, and have calculated a time-dependent effective mutation rate $U_e(t)$. We note that this idea that purifying

selection leads to a time-dependent mutation rate has been suggested by several recent studies (Woodhams 2006; O'Fallon 2010), and evidence for such time dependence has been presented in humans, fish, and birds (Ho *et al.* 2005; Penny 2005; Burridge *et al.* 2008; Weir and Schluter 2008). Our analysis shows the precise form of the time-dependent mutation rate we expect due to purifying selection, though it remains unclear whether this effect is responsible for the signatures found in recent data.

By combining our result for the time-dependent mutation rate with our time-dependent effective population size, we can in principle calculate any statistic of interest describing patterns of molecular evolution. If we treat mutations as a Poisson process, the probability that m mutations occur along n genealogical branches of length t, beginning at t_0, is given by

$$P_n(m|t,t_0) = \frac{\left[\int_{t_0}^{t} nU(t')dt'\right]^m}{m!} \exp\left[-\int_{t_0}^{t} nU(t')dt'\right]. \tag{4.15}$$

However, we note that this expression involves a subtle approximation. Although neutral mutations may be treated as a Poisson process with constant rate U_n, deleterious mutations are not strictly a nonhomogeneous Poisson process. This is because mutation rates at different times are not independent: the actual deleterious mutations are constrained by the fitness classes of individuals, such that if a mutation occurs at a particular time t, the probability of mutations at other times is constrained. Therefore, it is not strictly appropriate to use the Poisson approximation of Eq. (4.15). However, this approximation is closely related to the independent lineages approximation. For example, consider the ancestry of a single individual. Formally, the individual is drawn from a fitness class k that is Poisson distributed, and the total number of deleterious mutations in the ancestry of this individual must be exactly k. As a consequence, the number of deleterious mutations in the ancestry of a randomly-chosen individual is Poisson distributed with mean $\frac{U_d}{s}$. In contrast, in our expression for $U_e(t)$, we average over all classes from which this individual could have been sampled, and we treat deleterious mutations as a nonhomogeneous Poisson process at this rate. Thus, the number of deleterious mutations in the ancestry of

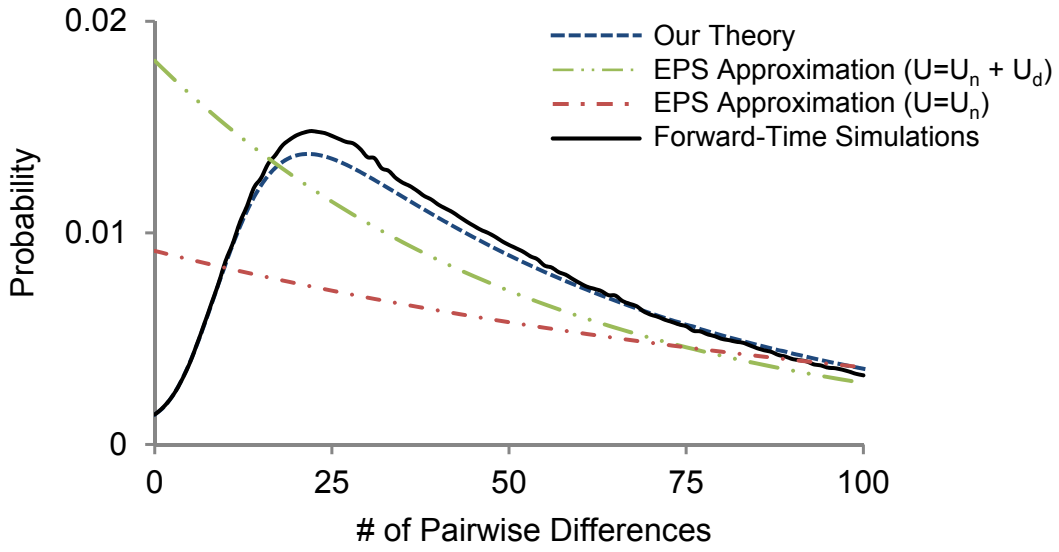

Figure 4.6: **Number of Pairwise Differences between Two Individuals,** for $U_d = 0.002$, $N = 10^5$, $U_n = U_d$, and $s = 10^{-3}$. We compare our theoretical result with forward-time simulations. For reference, we include the effective population size (EPS) approximation for both $U = U_n + U_d$ and $U = U_n$.

the individual is again Poisson distributed with mean $\int_0^\infty U_d e^{-st} dt = \frac{U_d}{s}$. This correspondence will no longer hold explicitly when tracing larger samples of individuals through the fitness distribution, because the ancestral histories are interdependent, and the formal class structure needs to be taken into account. However, provided the independent lineages approximation holds, we expect these errors to be small. To confirm the validity of this approximation, we can compare our theoretical result with forward-time simulations. For example, the distribution of the number of pairwise differences in a sample of two individuals is given by

$$P(\Pi = \pi) = \int_0^\infty \Psi_2(t) P_2(\pi|t) dt. \tag{4.16}$$

We compare this theoretical result with forward-time simulations in Fig. 4.6. More complicated statistics can be calculated in an analogous manner.

4.6 Applications

Our results demonstrate that patterns of molecular evolution in a population undergoing strong purifying selection are identical to those in a purely neutral population with the appropriate $N(t)$ and $U(t)$. This has the potential to aid in the analysis of populations experiencing strong purifying selection, by allowing us to describe such populations using an entirely neutral framework. Most importantly, it implies that preexisting neutral methods of population genetic inference can be used to estimate selection pressures, simply by incorporating the appropriate time-dependent population size and mutation rate. This avoids the difficulties inherent in full methods of inference using models that explicitly include selection, such as the need to identify each mutation as deleterious or neutral, and with summing over the possible combinations of fitness classes.

To show that this correspondence between purely neutral methods and models incorporating selection is indeed accurate, we ran a set of neutral coalescent simulations for a sample of size 15. These simulations assume that the population is entirely neutral, but with the appropriate time-varying size and mutation rate, $N_e(t)$ and $U_e(t)$, which our analysis has shown corresponds to a particular selected situation. In Fig. 4.7, we compare these results with forward-time simulations of a population undergoing strong purifying selection. In the figure, we show comparisons of the average number of pairwise differences, the total number of segregating sites, Tajima's D, and Fu and Li's D, for a sample of size 15. For comparison, we also show the EPS approximation result. We see that the neutral model with the appropriate $N_e(t)$ and $U_e(t)$ accurately captures a significant distortion due to selection in the shape of the genealogies. The agreement is good but not perfect — for example, as seen in Fig. 4.3, our formula for $N_e(t)$ slightly underestimates the long-term $N_e(t)$, such that our neutral coalescent simulations underestimate the branch lengths in the distant past, leading to overestimates of Tajima's D and Fu and Li's D. However, these systematic errors are small, and our analysis still accurately captures the general distortion in the distribution of these statistics.

These results demonstrate that preexisting neutral coalescent-based methods of

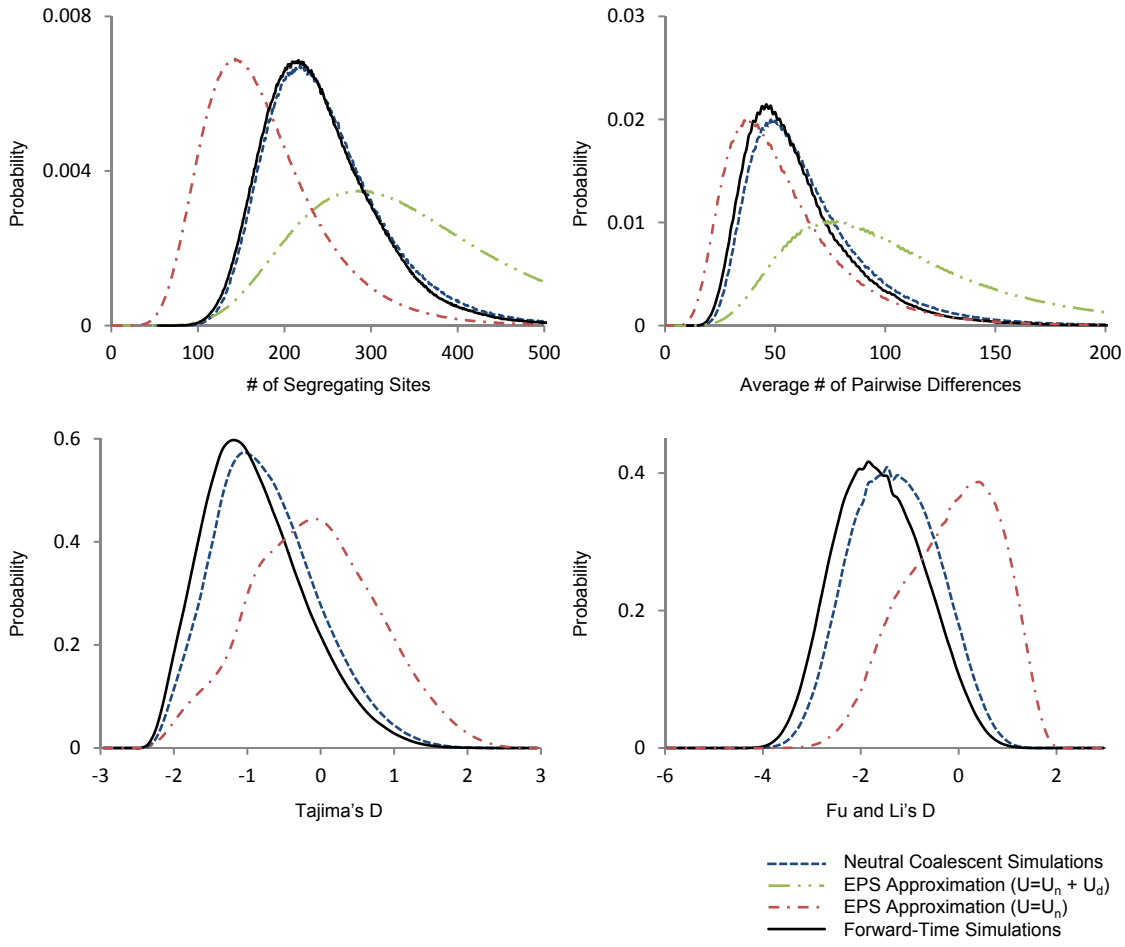

Figure 4.7: **Statistics for a Sample of Size** 15**,** for $U_d = 0.002$, $N = 10^5$, $U_n = U_d$, and $s = 10^{-3}$. We compare neutral coalescent simulations, using our $N(t)$ and $U(t)$, with forward-time simulations under purifying selection. For reference, we include the effective population size (EPS) approximation.

inference can be used for populations undergoing strong purifying selection by using the appropriate $N_e(t)$ and $U_e(t)$. A recent study by O'Fallon (2010) used a similar approach, in which the author incorporated a time-varying apparent mutation rate into likelihood calculations for genealogical inference in LAMARC. This method was then applied to data from the mitochondrion of the North Atlantic *Cyamus ovalis*. In this study, the decline in the apparent mutation rate was described by the ad-hoc function

$$\lambda(t) = 1 - \alpha(1 - e^{-\beta t}). \tag{4.17}$$

In comparison, our analysis shows that this function should be

$$\frac{U(t)}{U(0)} = 1 - \frac{U_d}{U_n + U_d}(1 - e^{-st}). \tag{4.18}$$

Thus, our analysis demonstrates that the function proposed by O'Fallon (2010) has the correct form, and allows us to identify the parameters he infers from his models with the actual selection pressures and mutation rates, provided the population is evolving in the strong selection regime.

O'Fallon (2010) compared his 'purifying rate' model with forward-time simulations, and observed a significant improvement over neutral models in inferring the time to the most recent common ancestor. However, his method could not account for the fact that selection is also expected to distort the genealogies, in addition to creating a time-dependent mutation rate. Our results provide a method to overcome this difficulty — we simply incorporate the appropriate time-varying population size $N_e(t)$ that corresponds to the same selection pressures as assumed in the time-varying mutation rate $U_e(t)$. By extending the analysis of O'Fallon (2010) to also include this $N_e(t)$, it is possible to perform full-scale inference on populations undergoing strong purifying selection, simultaneously accounting for both the non-uniform distribution of mutations, as well as the distortions in the shape of genealogies. This has the potential to significantly improve methods of dating and inference for such populations.

In addition to full-scale inference methods, our results also have significant implications for data from recent studies investigating apparent time-dependence in the

molecular clock (Ho *et al.* 2005; Burridge *et al.* 2008; Weir and Schluter 2008). These studies rely on analyzing sequences where divergence times can be estimated through geographical or fossil evidence. This data can then be used to estimate a mutation rate at different calibration times. The simplest method is the following: in neutrally evolving populations, the expected number of pairwise differences between two individuals is equal to two times the mutation rate times the coalescence time. By comparing the measured number of pairwise differences with the estimated divergence time, a mutation rate can be inferred. Several recent studies have shown that the mutation rate estimated using this method depends on the time at which coalescence occurs, with recent coalescence events implying a larger mutation rate than more ancient coalescence events. In other words, the mutation rate is apparently declining into the past (referred to as the "J-shaped curve") (Woodhams 2006; O'Fallon 2010; Ho *et al.* 2011).

Our analysis provides a way to determine whether these observations can be explained by the action of purifying selection, and to estimate the selection pressures involved. In our model, the expected number of pairwise differences divided by the coalescence time is $\mu(t) = \frac{\int_0^t U(t')dt'}{t} = U_n + U_d\left(\frac{1-e^{-st}}{st}\right)$, which we refer to as the "time-averaged apparent mutation rate". For very short times, $\mu(t) \to U_n + U_d$, indicating that selection has not yet had time to remove recent deleterious mutations from the population. However, at long times, $\mu(t)$ falls off to U_n, indicating that ancient deleterious mutations have been removed. The transition between these extremes is decreasing with rate given by the selection coefficient, s.

Our result for $\mu(t)$ provides a way to determine whether purifying selection is a likely explanation for the observed time-dependence in recent studies, and to directly estimate the neutral mutation rate, the deleterious mutation rate, and the selection coefficient, provided the population is evolving under strong purifying selection. For example, Burridge *et al.* (2008) studied the divergence rate in New Zealand freshwater fish as a function of time and found evidence for a time-dependent mutation rate. The authors analyzed samples of fish mtDNA from isolated geographical locations that were once connected, and estimated the time of the isolation events. They then used the isolation model of Wakeley and Hey (1997) to infer a divergence time

(scaled by the mutation rate). By comparing this with their estimates of the isolation events, the authors were able to infer mutation rates for isolation times ranging from $0.007 - 5.0$ Myr. They found that the resulting mutation rates were elevated in the recent past, on a time-scale of approximately 200 kyr. Specifically, they fit an exponential decay curve to data from galaxiidae, yielding a rate of change per site per million years of $0.02 + 0.04e^{-5.3t}$. If we compare this result with our $\mu(t)$, this would imply a per-site per-generation neutral mutation rate of 2×10^{-8} and a per-site per-generation deleterious mutation rate of 4×10^{-8}. Our function $\mu(t)$ decays more slowly than exponentially, implying a selection coefficient of approximately two to three times the fitted exponential decay rate, or about 10^{-5}. We note, however, that the error bars on the short-term data points are large, such that the 95% confidence intervals for the selection coefficient and deleterious mutation rate are high.

Importantly, we note that several other explanations have been proposed to explain the time-dependence of the mutation rate and may significantly contribute to the observed rate in this case. However, our result provides an informative way to interpret this data and suggests that purifying selection is a plausible explanation for the observed results. In order to test this hypothesis in detail, it is now possible to use our formula for $U_e(t)$ to perform a similar inference test to that performed in Burridge *et al.* (2008), without assuming a constant, fixed mutation rate. This would provide us with a method to estimate both the neutral and deleterious mutation rates, as well as the selection coefficient. One of the main benefits of this method is that the inferred mutation rates and selection coefficient in turn imply a particular $N_e(t)$. Thus, if the observed time-dependence is a result of purifying selection, we expect the population to be described by the corresponding $N_e(t)$, whereas a different population size may be expected if the time-dependence is a consequence of other effects (such as an actually varying mutation rate).

Interestingly, as an example of this possibility, another study by Zemlak *et al.* (2010) looked at the effects of historical climate factors in the Patagonian fish *Galaxias maculatus*. In their study, they estimated the effective population size as a function of time using a Bayesian skyline model, and similarly found a decay over a time-period of $200 - 500$ kyr, with an approximately 100-fold decay between the instantaneous

effective population size and the long-term effective population size. Although this result may be explained by climate effects that occurred on a similar timescale, this behavior is consistent with a population undergoing purifying selection with $\frac{U_d}{s} \approx \ln 100 \approx 4.6$ and selection coefficient of $s \approx \frac{1}{\text{timescale}} \sim 5 \times 10^{-6}$.

We note that our results hold only within the strong selection regime, when $Nse^{-U_d/s} \gg 1$. Thus, it is unclear whether our results will accurately describe these specific data sets. In each case, we estimate s of order 10^{-5}. This then requires a long-term effective population size of at least $\approx 10^5$ in order for the condition $Nse^{-U_d/s} \gg 1$ to hold. Thus, it is essential to jointly estimate the parameters using full-scale inference methods along the lines of O'Fallon (2010), as described above, in order to assess whether our results can be used to describe a particular data set. This is an interesting topic for future work.

In general, we caution that our results hold only within the strong selection regime, when $Nse^{-U_d/s} \gg 1$. Furthermore, our results hold only in non-recombining regions of the genome. This lack of recombination can potentially imply a large number of linked selected sites, which may in turn imply a large $\frac{U_d}{s}$. Therefore, it is important to ensure that the strong selection condition is met in order to avoid misleading results.

4.7 Conclusion

In summary, we have calculated a time-dependent mutation rate and a time-dependent effective population size that can be used to describe a population undergoing purifying selection. Our expression for $N_e(t)$ shows that recent genealogical branches are increased in length relative to older branches, leading to an increase in rare mutations relative to an undistorted model. This agrees with the qualitative conclusions of previous work (Williamson and Orive 2002; O'Fallon *et al.* 2010; Seger *et al.* 2010). Our expression for $U_e(t)$ shows that in addition to this effect, deleterious mutations are not uniformly distributed across the branches, and instead are biased even further towards the more recent branches.

We note that our method breaks down for weak selection in small populations, since both the steady state approximation and the independent lineage approxima-

tion break down. Within the parameter regime we consider, $N_e(t)$ is the same for any sample size, such that the genealogical trees are topologically neutral. However, as selection becomes weaker there is no longer any single $N_e(t)$ that applies to all samples. This implies that in addition to causing distortions in branch lengths, purifying selection also distorts the distribution of genealogical topologies. These topological distortions offer potential statistical power to distinguish purifying selection from demographic effects in patterns of molecular evolution. Our analysis has pointed to the parameter regimes in which we can expect these topological distortions to exist. Developing a simple analytical description of the nature of these topological distortions remains an interesting and important topic for future work.

4.8 Acknowledgements

We thank John Wakeley, Aleksandra Walczak, Joshua Plotkin, and Benjamin Good for many useful discussions. This work was supported by the James S. McDonnell Foundation, the Harvard Milton Fund, and the Alfred P. Sloan Foundation. LEN is supported by the Department of Defense through the National Defense Science and Engineering Graduate Fellowship Program. Simulations were run on the Odyssey cluster supported by the FAS Sciences Division Research Computing Group at Harvard University.

4.9 Appendix A: Calculation of the Ancestral Fitness Distribution

In this Appendix, we calculate the ancestral fitness distribution $P(k_i \to k_f|t)$. We have from the main text that

$$P(k_i \to k_f|t) = \frac{1}{sk_f} \int \delta(t - \sum t_j) \prod_{j=0}^{j=k_i-k_f} s(k_i - j)e^{-s(k_i-j)t_j} dt_j$$

In general, the convolution of n exponential distributions with parameters $\{\lambda_0, \lambda_1 \ldots \lambda_n\}$ is

$$\sum_{i=0}^{n} \lambda_i e^{-\lambda_i t} \prod_{j=0,\neq i}^{n} \frac{\lambda_j}{\lambda_j - \lambda_i},$$

as described by Wakeley (2009). Therefore, we have

$$P(k_i \to k_f|t) = \sum_{i=0}^{k_i-k_f} e^{-s(k_i-i)t} \left(\frac{\prod_{j=0}^{k_i-k_f-1} k_i - j}{\prod_{j=0,\neq i}^{k_i-k_f} i - j} \right).$$

We can use the fact that

$$\prod_{j=0}^{k_i-k_f-1} (k_i - j) = (k_i)(k_i - 1)\ldots(k_f + 1) = \frac{k_i!}{k_f!}$$

and

$$\prod_{j=0,\neq i}^{k_i-k_f} (i - j) = i(i-1)\ldots(1)(-1)(-2)\ldots(i - k_i + k_f) = i!(k_i - k_f - i)!(-1)^{k_i-k_f-i}$$

to write

$$P(k_i \to k_f|t) = \sum_{i=0}^{k_i-k_f} (-1)^{k_i-k_f-i} e^{-s(k_i-i)t} \binom{k_i - k_f}{i} \binom{k_i}{k_f} \tag{4.19}$$

$$P(k_i \to k_f|t) = e^{-sk_i t}(-1)^{k_i-k_f} \binom{k_i}{k_f} \sum_{i=0}^{k_i-k_f} (-e^{st})^i \binom{k_i - k_f}{i}. \tag{4.20}$$

Using the binomial equation:

$$(1 + x)^n = \sum_{i=0}^{n} x^i \binom{n}{i},$$

and identifying $x = -e^{st}$ and $n = k_i - k_f$, this becomes

$$P(k_i \to k_f|t) = e^{-sk_it}(e^{st} - 1)^{k_i - k_f}\binom{k_i}{k_f}, \tag{4.21}$$

as claimed in the main text.

Chapter 5

Distortions in Genealogies due to Purifying Selection and Recombination

Purifying selection at many linked sites alters patterns of molecular evolution, reducing overall diversity and distorting the shapes of genealogies. Recombination attenuates these effects, however purifying selection can significantly distort genealogies even for substantial recombination rates. Here, we show that when selection and/or recombination are sufficiently strong, the genealogy at any single site can be described by a time-dependent effective population size, $N_e(t)$, which has a simple analytic form. Our results illustrate how recombination reduces distortions in genealogies, and allow us to quantitatively describe the shapes of genealogies in the presence of strong purifying selection and recombination. We also analyze the effects of a distribution of selection coefficients across the genome.

5.1 Introduction

Purifying selection acts to remove deleterious mutations, and variation linked to these mutations, as they continually arise in a population. This leads to reduced genetic diversity at both selected sites and linked neutral sites. This effect, known as background selection, can significantly impact the patterns of diversity evident in sequence data. In recent years, substantial empirical evidence has arisen suggesting that these effects may be pervasive in humans and other organisms (McVicker *et al.* (2009); Lohmueller *et al.* (2011); Haddrill *et al.* (2010); see Charlesworth (2012) for review).

When selection is very strong, these effects are well understood. Deleterious mutations are purged almost immediately, leading to an overall reduction in diversity to the level expected in a neutrally evolving population with a reduced population size N_e (Hudson and Kaplan 1994, 1995a; Nordborg *et al.* 1996; Charlesworth 1994). The size of this reduction depends upon the recombination rate, which determines the extent to which each site is linked to potentially deleterious mutations. This effect is captured by a simple analytic formula showing how N_e depends upon mutation rates, selection strengths, and recombination rates, and has been widely used to interpret patterns of molecular evolution (Hudson and Kaplan 1995b; Charlesworth 2013).

However, it has long been recognized that in addition to overall reductions in diversity, purifying selection also distorts the shapes of genealogies (Charlesworth *et al.* 1993, 1995; Zeng and Charlesworth 2011). These distortions arise because purifying selection does not act instantaneously, and hence deleterious mutations can persist transiently in the population, lengthening coalescence times in the recent past relative to those in the distant past (Barton and Etheridge 2004; Williamson and Orive 2002). A number of recent studies have addressed these distortions in the completely nonrecombining (asexual) case (Seger *et al.* (2010); O'Fallon *et al.* (2010); Walczak *et al.* (2012); Nicolaisen and Desai (2012); see Charlesworth (2013) for review). However, very little is quantitatively known about the shape and magnitude of these distortions in the presence of recombination.

To address this question, Zeng and Charlesworth (2011) recently developed a structured coalescent algorithm to simulate genealogies in the presence of purifying selection with recombination, analogous to the asexual structured coalescent (Hudson and Kaplan 1994). They used this method to analyze distortions in statistics such as the mean coalescence time and the ratio of external branch length to total branch length. This approach makes it possible to rapidly simulate any statistic describing the shape of genealogies, including those involving multiple sites (e.g. the correlation in coalescence times at two sites). Although this approach is only valid for sufficiently strong selection, it has led to many novel conclusions and offers great promise as a useful practical tool.

However, despite these advantages, one of the main difficulties of simulation-based methods is that they cannot provide simple analytical insight into how distortions in genealogies depend upon the relevant parameters, and it is thus difficult to incorporate results into practical inference or estimation methods. In this paper, we show that the distortions in genealogies may be described by a neutral population with a time-dependent effective population size $N_e(t)$, and we compute a simple analytical formula describing how this $N_e(t)$ depends upon mutation rates, selection strengths, and recombination rates. Our approach is closely related to our earlier analysis of the effects of purifying selection in completely nonrecombining regions (Nicolaisen and Desai 2012), and many of our results are closely analogous, demonstrating that many of the effects of selection in asexual populations remain qualitatively similar in the presence of substantial recombination.

Our analysis is limited to describing genealogies at a single site, and is only valid provided purifying selection and recombination are sufficiently strong. Thus, it is unable to describe the topological distortions in genealogies which begin to appear as selection and recombination become weaker. However, despite these limitations, our results show that the effects of strong purifying selection and recombination may be described by a time-varying population size, and explicitly describe the analytical dependence of this time-varying population size on the underlying parameters. This result can therefore be directly incorporated into pre-existing neutral methods of inference and estimation in a time-varying population to describe or infer the effects

of selection and recombination. As we will see, it is also straightforward to incorporate the effects of variation in selection strengths across sites into our analysis.

We begin in the next section by describing our model, which is closely related to earlier studies of background selection with recombination (Hudson and Kaplan 1994, 1995b,a; Nordborg *et al.* 1996; Charlesworth *et al.* 1993). We focus on a single focal site in a randomly-sampled individual, and we trace that individual's ancestral history backwards in time. We calculate the probability that a single linked site carries a deleterious mutation as a function of time in the past. In general, selection acts against individuals carrying deleterious mutations, such that the probability an ancestor carried such a mutation decreases as we look further into the past. We refer to this probability as the "ancestral fitness distribution," and we use this to calculate the probability that two individuals contain the same set of deleterious mutations across all linked sites. This calculation will rely upon the key assumption that we may treat the ancestral fitness distribution at each site as independent (see Supplemental Information B.1). Finally, we use this to calculate the probability of coalescence over time and thus, $N_e(t)$.

5.2 Analysis

We consider a haploid population of N individuals. For simplicity, throughout most of our analysis we will assume that each site experiences deleterious mutations at rate μ to an allele carrying selective cost s. In a later section below, we show how our results can be straightforwardly generalized to the case where each site has an arbitrary mutation rate and selection coefficient. Throughout, we assume that there is no epistasis, and that fitnesses combine multiplicatively, so that an individual with k deleterious mutations has fitness $(1-s)^k$. We note that this haploid model is analogous to that of Zeng and Charlesworth (2011) and is closely related to earlier diploid models of purifying selection and recombination (Hudson and Kaplan 1994, 1995b,a; Nordborg *et al.* 1996; Charlesworth *et al.* 1993) – resulting equations can be compared with simple modifications.

5.2.1 The Ancestral Fitness Distribution

We consider a single individual, randomly sampled from the population. We wish to trace the ancestral history of that individual at a single focal site backwards-in-time, in order to calculate the probability that the ancestor carries a mutation at a particular linked site (referred to as the "index site"), as a function of time in the past.

In the present, we know that the probability an individual carries a deleterious mutation at a particular site is given by the classical mutation-selection balance result, $p_{mut} = \mu/s$ (Kimura and Maruyama 1966; Haigh 1978). However, there are two types of events that may occur in the ancestral history. First, a deleterious mutation may occur, which will move the ancestor from the mutant state into the non-mutant state. This will occur at rate $\frac{\mu N(1-p_{mut})}{Np_{mut}} \approx s$, where we have neglected terms of higher-order in μ/s. Second, a recombination event may occur between the index and focal sites at rate $r(x_i, x_f)$, where x_i denotes the index site and x_f denotes the focal site. When a recombination event separates the focal site from the index site, the ancestor at the index site is randomly chosen from the population (Figure 1). Thus, the ancestor will carry a mutation with probability $p_{mut} = \mu/s$. Writing this out, we have that

$$\frac{dP_{mut}(t)}{dt} = -(s + r(x_i, x_f))P_{mut}(t) + r(x_i, x_f)\mu/s,$$

where $P_{mut}(t)$ is the probability the ancestral lineage carries a deleterious mutation at the index site at time t. We note that we have neglected the effects of back mutations, which introduce terms of higher-order in μ/s. However, it is straightforward to include these terms (see Supplemental Information B.2). Solving the above differential equation, we have that the ancestral fitness distribution is simply

$$P_{mut}(x_i, x_f, t) = \frac{\mu}{s}\left(\frac{r(x_i, x_f)}{r(x_i, x_f) + s} + \frac{s}{r(x_i, x_f) + s}e^{-r(x_i,x_f)t - st}\right). \tag{5.1}$$

We note that Eq. (5.1) allows for recombination rates to vary in any arbitrary way across the genome. However, to illustrate our main results, it is often helpful to make

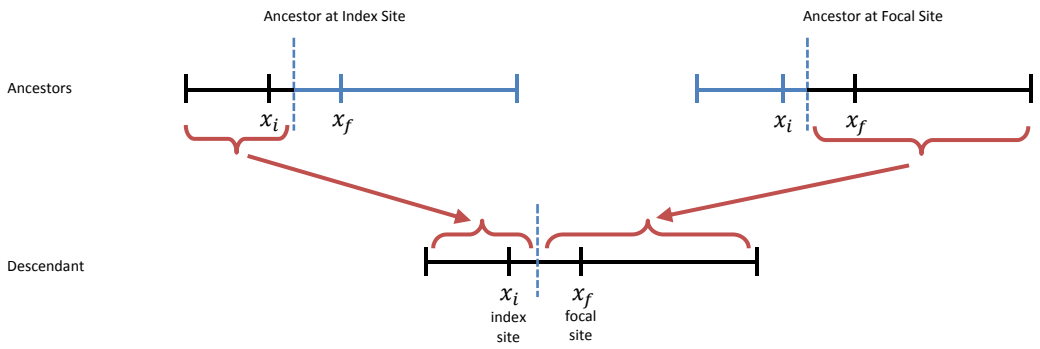

Figure 5.1: **A Recombination Event in an Ancestral Lineage:** Each parent has part of the genome ancestral to the descendant lineage (black); these have an ancestral fitness distribution unaffected by the recombination event. The non-ancestral parts of the parental genomes (blue) are effectively sampled at random from the population, and hence have a fitness distribution reflective of the steady-state mutation-selection balance. Throughout this paper, we focus only on the genealogies at a single focal site x_f, and hence only track the parent with the ancestral sequence at this site (right branch in this figure).

the simplifying assumption that recombination occurs at constant per-site rate r. In this case, denoting $x \equiv |x_f - x_i|$, Eq. (5.1) reduces to

$$P_{mut}(x,t) = \frac{\mu}{s}\left(\frac{rx}{rx+s} + \frac{s}{rx+s}e^{-rxt-st}\right). \tag{5.2}$$

For clarity we use this simpler expression throughout most of our subsequent analysis, but we note that our results can all be generalized to account for the effects of variation in recombination rates by replacing Eq. (5.2) with Eq. (5.1) throughout.

Intuitively, Eq. (5.2) reflects the fact that sites far from the focal site will typically recombine away more frequently, and hence have an ancestral fitness distribution that is closer to that given by the steady state mutation-selection balance, μ/s. By contrast, sites very close to the focal site are unlikely to recombine away, and the ancestral lineage at these sites is biased to be more fit than average. The distance $L^* = \frac{s}{r}$ is the boundary between these two regimes, and represents the natural length scale on which the effects of selection are diminished. On distances small compared

to L^*, the ancestral fitness distribution is comparable to the nonrecombining case, since recombination occurs very infrequently compared to selection. By contrast, on distances large compared to L^*, recombination occurs so rapidly that selection does not have time to remove mutations from the ancestral lineage before it is reset by recombination.

We note that, in analogy with earlier work on background selection with recombination, as well as the structured coalescent method of Zeng and Charlesworth (2011), this derivation assumes that all sites may be treated deterministically. This approximation will be reasonable provided no individual lineage becomes a significant fraction of the population (or, analogously, if the typical time-scale on which deleterious mutations are removed form the population is short relative to the population size), which holds when $Nse^{-\frac{\mu L}{s+rL/2}} \gg 1$. When this approximation breaks down, mutant lineages may grow to a significant fraction of the total population, and our results will no longer accurately capture the ancestral fitness distribution.

We compare our theoretical result in Eq. (5.2) with forward-time simulations in Fig. 5.2. We compare the ancestral fitness distribution at five different ancestral timepoints, in two different parameter regimes. We see that for strong-selection/recombination, Eq. (5.2) accurately describes the ancestral fitness distribution at each timepoint, as predicted. By contrast, when selection and/or recombination become weaker, we see that our Eq. (5.2) systematically overestimates the ancestral fitness. Thus, as expected, our results become less accurate as $N_e s$ becomes small and the deterministic approximation breaks down. This is a consequence of fluctuations: as the strong-selection/recombination condition breaks down, mutant lineages may occasionally grow to a substantial fraction of the population. This will systematically bias mean mutation probabilities to higher values. As we will later see, despite these events, we will still be able to accurately predict the shape of genealogies as the strong-selection/recombination condition begins to be violated. However, strong violations of this condition will cause our results to break down by introducing additional distortions which we are not able to address.

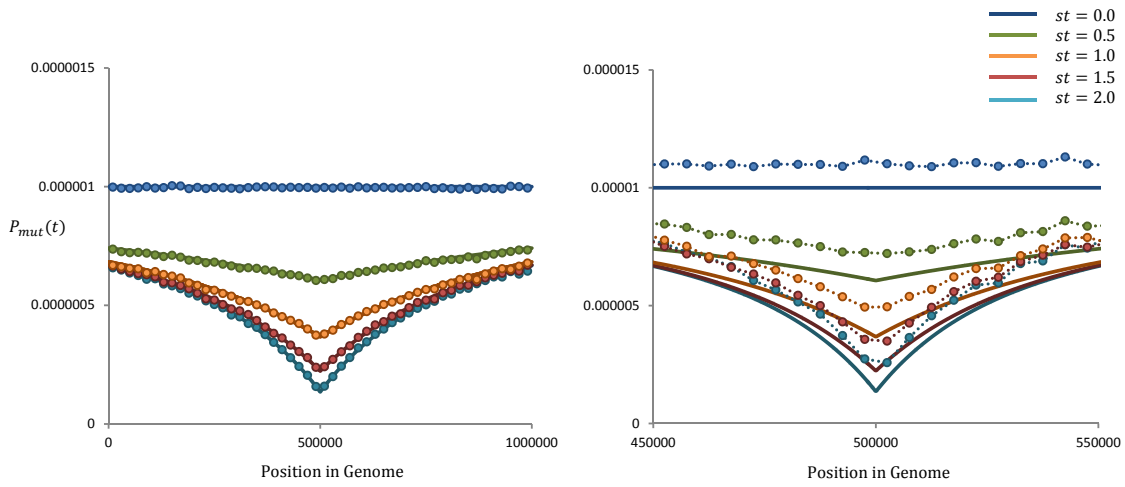

Figure 5.2: **The Ancestral Fitness Distribution as a Function of Position:** In both figures, $\mu = 10^{-8}$, $N = 10^4$, and $r = 4 * 10^{-8}$. The focal site is located at the center of a genome of length $L = 10^6$. Our theoretical results are shown as solid lines, while the simulations are represented with circles. In the first figure, $s = 10^{-2}$, such that $Nse^{-U_d/(s+R/2)} = 71.7$. In the second figure, $s = 10^{-3}$, such that $Nse^{-U_d/(s+R/2)} = 6.2$. We see that as $N_e s$ becomes smaller, the deterministic approximation begins to break down, and fluctuations in the population occasionally allow lineages carrying deleterious mutations to become a large fraction of the population. This, in turn, leads to less-fit ancestral lineages than predicted by our theoretical results.

We note that in deriving Eq. (5.2), we have assumed a continuous approximation which requires the per-generation rate of recombination to be small ($rx \ll 1$). This will only be strictly valid if the total genome-wide recombination rate is small, $rL \ll 1$. However, in practice, our result will still be valid even when $rL > 1$, since only sites close to the focal site contribute to the effects of selection on genealogies. To be specific, only sites within $x \lesssim L^*$ of the focal site are significantly affected by selection, so Eq. (5.2) will in fact be roughly valid provided only that $rL^* \approx s \ll 1$, which we generally expect to hold. In a similar vein, instead of approximating $r(x) = rx$, it would be more appropriate to use a mapping function such as the Haldane formula, $r(x) = \frac{1-e^{-2rx}}{2}$ (Haldane 1919). However, this is also roughly equivalent to our approximation within the relevant range, provided only that $s \ll 1$. This was earlier

noted by NORDBORG *et al.* (1996), who observed that the choice of mapping function was not significant since only closely linked sites contribute to the effects of selection.

We have derived an ancestral fitness distribution, which captures the probability that the ancestor of an individual will have a mutation at a particular site in the past. We now use this ancestral fitness distribution to calculate the probability that two or more individuals share the same set of mutations, which will enable us to calculate the probability of coalescence as a function of time.

5.2.2 The Effective Population Size

When two ancestral lineages have the same set of mutations across all sites (i.e. they are in the same "configuration"), they coalesce with per-generation probability $1/N_{config}$, where N_{config} is the total number of individuals in that configuration. Thus the probability that two arbitrary ancestral lineages coalesce a time t in the past is the probability they exist in the same configuration divided by the number of individuals in that configuration,

$$P_c(t) = \frac{1}{N_e(t)} = \sum_{\text{configurations}} \frac{P_{\text{config}}(t)^2}{N_{\text{config}}}.$$

As in earlier work on background selection with recombination, provided the deterministic approximation holds and $\mu/s \ll 1$, we may treat each site as independent (see Supplemental Information B.1 for further details about this approximation). We then have:

$$\begin{aligned}
\frac{1}{N_e(t)} &= \frac{1}{N}\prod_{\text{sites}}\left[\frac{P_{\text{mut}}(x,t)^2}{P_{\text{mut}}(x,0)} + \frac{(1-P_{\text{mut}}(x,t))^2}{1-P_{\text{mut}}(x,0)}\right] \\
&\approx \frac{1}{N}\prod_{\text{sites}}\left[1 + \frac{\mu}{s}\left(\frac{s}{rx+s}(1-e^{-rxt-st})\right)^2\right] \\
&\approx \frac{1}{N}e^{\frac{\mu}{s}\sum\limits_{\text{sites}}\left(\frac{s}{rx+s}(1-e^{-rxt-st})\right)^2},
\end{aligned}$$

where we have neglected terms of order $\frac{\mu^2}{s^2}$ or higher. As before, it is possible to keep this entirely general, allowing the focal site to be at any position along a genome of arbitrary length. However, to illustrate our results, we will assume that the focal

site is located at the center of a genome of length L. Approximating the sum as an integral, this becomes:

$$N_e(t) \approx N e^{-\frac{2\mu}{s}\int_0^{\frac{L}{2}}\left(\frac{s}{rx+s}(1-e^{-rxt-st})\right)^2 dx}.$$

Carrying out the integral, we find

$$N_e(t) = N \exp\left[-\frac{2U_d}{R}\Big((1-e^{-st})^2 - \frac{s}{s+R/2}(1-e^{-st-Rt/2})^2 + 2st\big(\Gamma\left[0, st, 2st\right] - \Gamma\left[0, Rt/2+st, Rt+2st\right]\big)\Big)\right], \tag{5.3}$$

where we have defined $rL \equiv R$ and $\mu L \equiv U_d$. Although we derived this result by considering a sample of size two, we arrive at the same $N_e(t)$ for arbitrary sample sizes provided we may treat lineages as independent and exchangeable. This assumption holds provided the typical time-scale for backwards-in-time mutation events is short relative to the typical coalescence time (e.g. when $Nse^{-U_d/(s+R/2)} \gg \binom{n}{2}$, where n is the sample size). We note that this condition becomes more restrictive for larger samples, but provided it holds the coalescence probabilities are described by the $N_e(t)$ given above.

We see from Eq. (5.3) that in the recent past, the effective population size is simply $N_e(0) = N$. However, as time recedes into the past, the ancestral lineages become biased toward more fit configurations, and are correspondingly more likely to coexist in the same configuration concurrently. This implies that the rate of coalescence increases with time. As $t \to \infty$, our results reduce to $N_e(\infty) \to Ne^{-U_d/(s+R/2)}$, the haploid version of the original background selection result.

We compare our result for $N_e(t)$ in Eq. (5.3) with forward-time simulations in Fig. 5.3, as a function of genome size and recombination rate. Fig. 5.3 illustrates the significant period of transition from the larger N_e in the recent past, prior to reaching the long-term result, which results in distortions in the shapes of genealogies. The agreement is generally good, but we note that for smaller recombination rates our analysis systematically underestimates $N_e(t)$. This is a consequence of the deterministic approximation breaking down as the recombination rate decreases, leading to strong fluctuations in the population.

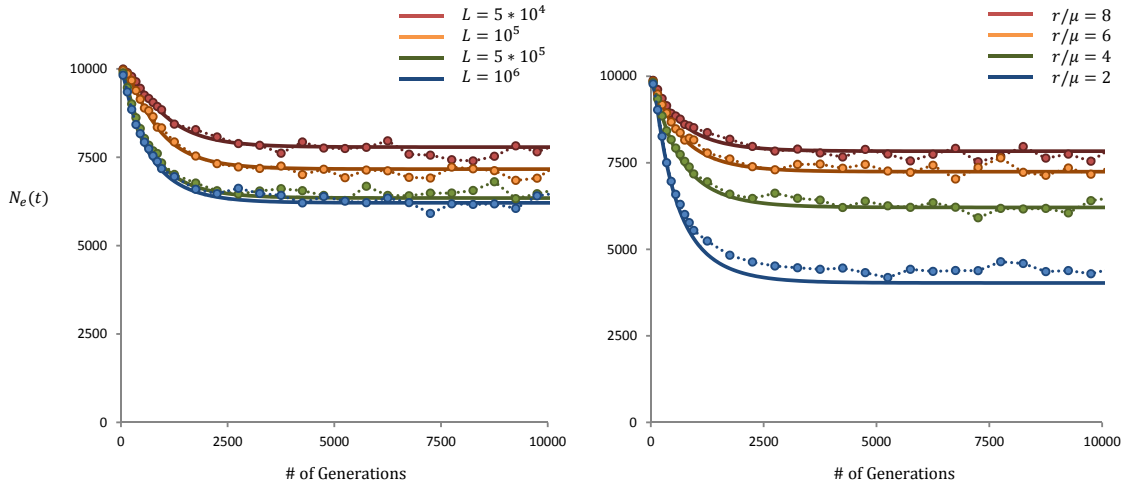

Figure 5.3: **Effective Population Size as a Function of Time:** In both figures, $\mu = 10^{-8}$, $s = 10^{-3}$, and $N = 10^4$. Our theoretical results are shown as solid lines, while the simulations are represented with circles. In the first figure, $r = 4 * 10^{-8}$ and L varies from $5 * 10^4$ to 10^6, such that $Nse^{-U_d/(s+R/2)}$ varies from 6.2 to 7.8. In the second figure, $L = 10^6$ and r/μ varies from 2 to 8, such that $Nse^{-U_d/(s+R/2)}$ varies from 4.0 to 7.8. The agreement is generally good, however, as seen in the second figure, as the recombination rate decreases and $N_e s$ falls off, the deterministic approximation begins to break down, and our results become less accurate.

Our results differ from the classical background selection results by incorporating the transient period during which deleterious alleles may segregate in the population prior to being removed. The time-scale of this transition period is, roughly, of order $1/s$ generations. In the deterministic regime, we have assumed that $N_e s \gg 1$, such that this transition period is, by definition, short relative to the typical coalescence times. However, despite this, as seen in Fig. 5.3, by accounting for this transition period we are able to capture a significant deviation from the classical result. This deviation is ultimately a primary source of the distortions we expect to see in genealogies, and thus our results are able to show how selection can lead to distortions in genealogies and in genealogical statistics, and how these effects depend upon the parameters involved. However, we note that our analysis is restricted to addressing the distortions that arise due to this transient period – when the deterministic ap-

proximation breaks down, fluctuations in the population become significant and lead to further distortions, including topological distortions, which our analysis is not able to capture.

5.2.3 Coalescence Times and other Single-Site Statistics

Our result for $N_e(t)$ leads immediately to an expression for the distribution of times to the next coalescence event in a sample of size n,

$$\Psi_n(t) = \frac{\binom{n}{2}}{N_e(t)} e^{-\int_0^t \frac{\binom{n}{2}}{N_e(t')} dt'}. \tag{5.4}$$

We compare this prediction to forward-time simulations for a sample of two individuals in Fig. 5.4; we see that there are significant distortions introduced by the time-dependence of $N_e(t)$, which lead to a nonzero peak in the distribution of coalescence times.

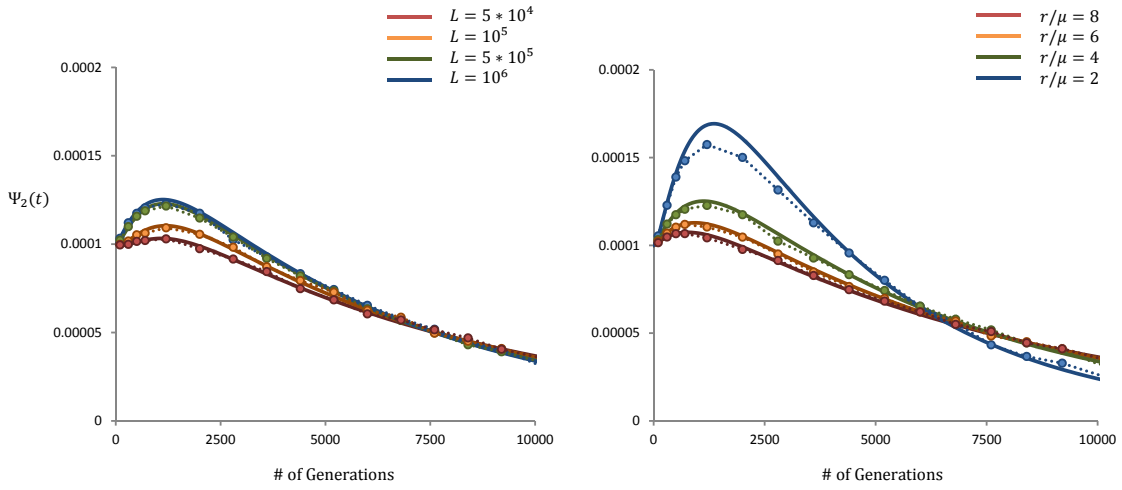

Figure 5.4: **Coalescence Probability as a Function of Time for a Sample of Size Two:** In both figures, $\mu = 10^{-8}$, $s = 10^{-3}$, and $N = 10^4$. Our theoretical results are shown as solid lines, while the simulations are represented with circles. In the first figure, $r = 4 * 10^{-8}$ and L varies from $5 * 10^4$ to 10^6, such that $Nse^{-U_d/(s+R/2)}$ varies from 6.2 to 7.8. In the second figure, $L = 10^6$ and r/μ varies from 2 to 8, such that $Nse^{-U_d/(s+R/2)}$ varies from 4.0 to 7.8.

Using the distributions of coalescence times, we can calculate various statistics describing genetic diversity. For example, the distribution of pairwise heterozygosity at a neutral focal site is given by

$$P(\Pi_{\text{neutral}} = \pi) \approx \int_0^\infty \frac{(2\mu t)^\pi}{\pi!} e^{-2\mu t} \Psi_2(t) dt. \tag{5.5}$$

In contrast, if the site is a selected site, then we have from Eq. (5.2) that the probability the ancestor carries a deleterious mutation is $P_{mut}(t) = \frac{\mu}{s} e^{-st}$. Backwards-in-time, an individual carrying such a mutation will undergo a deleterious mutation from the non-mutant state at rate $\frac{\mu N(1-p_{mut})}{N p_{mut}} \approx s$. Thus, the backwards-in-time mutation rate is simply μe^{-st} (Nicolaisen and Desai (2012)). Therefore, we can estimate that:

$$P(\Pi_{\text{deleterious}} = \pi) \approx \int_0^\infty \frac{(2\mu e^{-st})^\pi}{\pi!} e^{-2\mu e^{-st}} \Psi_2(t) dt. \tag{5.6}$$

Using similar logic, we can explicitly calculate more complex statistics using neutral methods incorporating a time-varying population size. For example, the mean time to the most recent common ancestor may be calculated using Eq. 4 of Austerlitz *et al.* (1997), and arbitrary moments of the total branch length may be calculated using Eq. 20 of Eriksson *et al.* (2010). Similarly, if the focal site is a neutral site, the site frequency spectrum may be calculated using Eq. 2 of Polanski and Kimmel (2003).

This approach is illustrated in Fig. 5.5. Here, we consider the total lengths of branches ancestral to i individuals in a sample of size 10 (normalized by the length of branches ancestral to 1 individual). We compare forward-time simulations (represented by circles) with our theoretical result. For comparison, we also show the result expected for any fixed effective population size. We see that there is a noticeable deviation from the neutral expectation, characterized by an excess of rare branch lengths relative to more common branches.

If the focal site is a neutral site, mutations occur uniformly along the branch lengths. Thus, our result in Fig. 5.5 would be directly analogous to the site frequency spectrum, and implies an excess of rare alleles relative to common ones. In contrast,

if the focal site is a selected site, deleterious mutations occur at a backwards-in-time rate of μe^{-st}. In this case, deleterious mutations will be further biased towards recent branches, leading to an even more pronounced excess of rare alleles relative to common ones. In order to understand the expected frequency spectrum in this case, or to calculate any other complicated genealogical statistic, we can implement purely neutral and nonrecombining coalescence simulations which account for the effects of selection and recombination simply by using the appropriate $N_e(t)$.

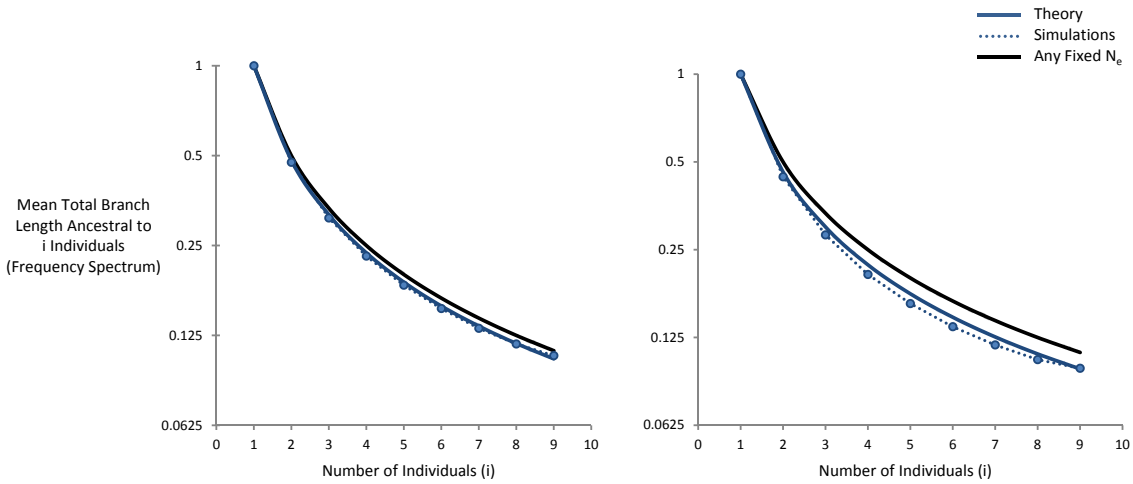

Figure 5.5: **Total Branch Length Ancestral to i Individuals for a Sample of Size Ten:** In both figures, $\mu = 10^{-8}$, $s = 10^{-3}$, $r/\mu = 4$, and $N = 10^4$. In the first figure, $L = 2.5 * 10^3$, such that $Nse^{-U_d/(s+R/2)} = 8.5$ (compared to $\binom{n}{2} = 45$). In the second figure, $L = 200 * 10^3$, such that $Nse^{-U_d/(s+R/2)} = 6.7$. Our theoretical result is shown as a solid line, while the simulations are represented with circles. The fixed effective population size result is shown as a solid black line.

5.2.4 Incorporating a Distribution of Fitness Effects

Our analysis can be easily extended to account for variation in recombination rates, mutation rates, and selection coefficients across the genome. Using the same logic described above, we find that in this more general case the time-dependent effective

population size is given by

$$N_e(t) \approx N \exp\left[-\sum_i \frac{2\mu(x_i)}{s(x_i)}\left(\frac{s(x_i)}{r(x_i,x_f)+s(x_i)}(1-e^{-r(x_i,x_f)t-s(x_i)t})\right)^2\right], \quad (5.7)$$

where $\mu(x_i)$ and $s(x_i)$ are the mutation rate and selection coefficient at site x_i respectively, and $r(x_i, x_f)$ is the total recombination rate between x_i and the focal site x_f. When $t \to \infty$, this reduces to the classical background selection result (Nordborg *et al.* 1996; Hudson and Kaplan 1995b; Charlesworth *et al.* 1996; Loewe and Charlesworth 2007).

A particularly interesting case is the situation where mutation and recombination rates are constant across the genome, but each selected site has a fitness effect drawn from some distribution $\rho(s)$. In this case, we find

$$N_e(t) \approx N \exp\left[-\int_0^{\frac{L}{2}}\int_0^{\infty} \frac{2\mu}{s}\left(\frac{s}{rx+s}(1-e^{-rxt-st})\right)^2 \rho(s)dsdx\right]. \quad (5.8)$$

We note that this result assumes that lineages can be treated deterministically, which requires that no lineage become a significant fraction of the total population, and thus $N_e s \gg 1$. Thus, we expect Eq. (5.8) to hold only when the bulk of the mutations are either in this regime, $N_e s_i \gg 1$, or nearly neutral $N_e s_i \ll 1$. Although our analysis is still reasonable when a small number of mutations exist in the intermediate regime, it requires that the bulk of the deviation from neutrality satisfies the strong-selection/recombination condition, such that these mutations of intermediate effect can be neglected. This issue was also addressed in earlier work considering a distribution of fitness effects (see e.g. Nordborg *et al.* (1996)).

Several recent studies have suggested that the distribution of deleterious fitness effects in humans and *Drosophila* may be characterized by a gamma distribution with shape parameter $\beta < 1$. For example, Keightley and Eyre-Walker (2007) estimated a shape parameter of $\beta \sim 0.2$ for human populations, and $\beta \sim 0.35$ for *Drosophila*. Motivated by these findings, we compare our theoretical results in Eq. (5.8) with forward-time simulations for two populations with gamma distributions of fitness effects, using shape parameters of 0.5 and 0.25, in Fig. 5.6. For reference, we also show the theoretical result expected under a single-s case, where s is the mean

fitness effect. We see that our results accurately characterize the time-dependence of the effective population size under a distribution of fitness effects.

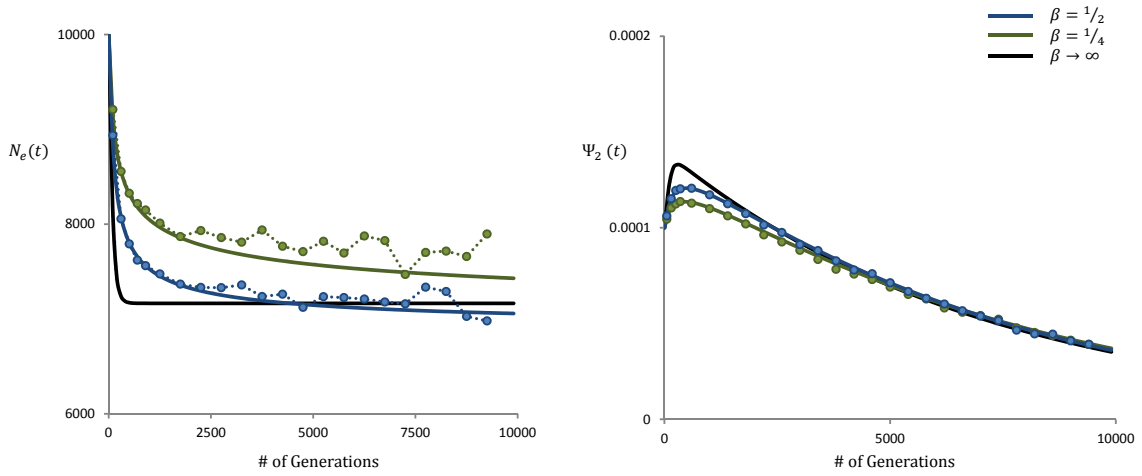

Figure 5.6: **Effective Population Size and Coalescence Times for a Distribution of Fitness Effects:** The blue curve shows a gamma distribution with shape parameter 1/2 and mean $s^* = 10^{-2}$. The green curve shows a gamma distribution with shape parameter 1/4 and mean $s^* = 10^{-2}$. The circles represent forward time simulations, while solid lines represent our theoretical results from Eq. (5.8). The black line shows our theoretical results in the single-s case with $s = s^*$. The parameters are: $N = 10^4$, $u = 10^{-8}$, $L = 10^6$, and $r/\mu = 4$.

5.2.5 Incorporating Temporal Variation in the Population Size

A key feature of our analysis is that, within this deterministic regime, the dynamics of ancestral lineages are independent of the population size. The implication of this is that the ancestral fitness distribution in Eq. (5.2) is independent of the population size, and thus Eq. (5.3) depends only upon an overall multiplication by N. This phenomenon was recently discussed in Zeng (2012), and was used as the basis for incorporating a changing population size into structured coalescent simulations. Similarly, we can incorporate a changing population size into our analysis, simply by replacing N with $N(t)$ in our Eq. (5.3).

We caution, however, that this framework implicitly assumes that the population remains in mutation-selection balance throughout its history, and is characterized by the same mutation rates, recombination rates, and selection coefficients throughout. If these other parameters are also changing with time, this will not hold. Similarly, the strong-selection/recombination condition must hold throughout the time-scale of coalescence.

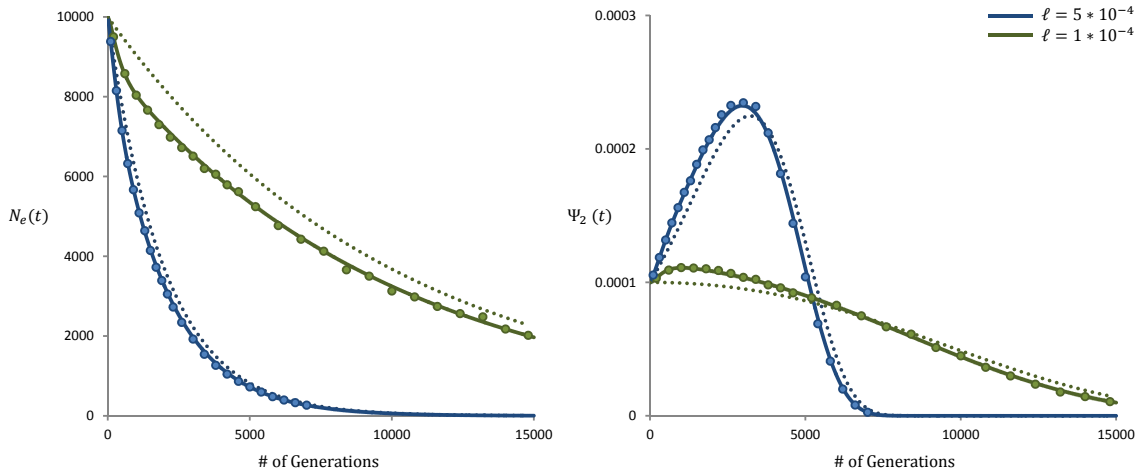

Figure 5.7: **Effective Population Size and Coalescence Times for an Exponentially-Growing Population:** The blue curves represent an exponentially growing population with growth rate $\ell = 5 * 10^{-4}$. The green curves represent an exponentially growing population with growth rate $\ell = 1 * 10^{-4}$. The circles represent forward time simulations, while solid lines represent our theoretical results from Eq. (5.3) with N replaced by $N(t) = Ne^{-\ell t}$. The dotted lines represent the predictions under a neutral model. The parameters are: $N = 10^4$, $s = 0.003$, $U_d = 0.0005$, and $R/U_d = 4$.

To illustrate this, in Fig. 5.7 we compare forward-time simulations for a population experiencing exponential growth $N(t) = Ne^{-\ell t}$ (with t measured backwards in time from the present) with our theoretical results from Eq. (5.3). For reference, we include the predicted theoretical result in the absence of selection.

5.2.6 Forward-time simulations

Our forward-time simulations are closely modeled off Zeng and Charlesworth (2011). Specifically, we simulate a haploid population of constant size N, with a genome of length L. Each generation, we introduce a Poisson-distributed number of new deleterious mutations uniformly throughout the population, with mean NU_d. We then simulate reproduction, introducing a Poisson-distributed number of recombination events uniformly throughout the population. For any non-recombinant offspring, one ancestor is chosen, weighted according to its fitness. For recombinant offspring, two ancestors are chosen, again weighted according to their fitnesses. The appropriate number of breakpoints are randomly chosen along the genome, and one of the two resulting genotypes is randomly selected as the offspring. These steps are repeated each generation.

We note that one key difference between our simulations and those of Zeng and Charlesworth (2011) is that we allow multiple recombination events to occur within a single individual. This makes it possible to consider the $L \to \infty$ limit, where multiple recombination events become common. We ran all simulations for a minimum of at least $10N$ generations to achieve equilibrium. When incorporating a distribution of fitness effects, we chose the fitness effect at each site according to the fitness distribution $\rho(s)$. In all figures, at least $10,000$ trials were completed for each parameter regime.

5.3 Discussion

Our analysis demonstrates that the effects of strong purifying selection and recombination can be summarized in terms of a time-dependent effective population size, $N_e(t)$. This $N_e(t)$ characterizes the distortions we expect to see in genealogies, and can be used as the basis for quick and efficient methods to analyze single-site statistics. It illustrates how these statistics depend upon parameters such as the selection coefficients, position in the genome, and variation in the recombination rate.

Our results extend earlier work that summarized the effects of purifying selection

and recombination by using a reduced but *constant* effective population size $N_e = N \exp\left[-\frac{U_d}{s+R/2}\right]$ (Charlesworth *et al.* 1993; Hudson and Kaplan 1995b). The simplicity of this earlier result has made it broadly useful in interpreting patterns in sequence data — for example, in determining whether background selection can be responsible for observed relationships between diversity and local recombination rates in humans and *Drosophila* (Hudson and Kaplan 1995b). Our analysis represents the simplest analytical extension of this earlier work that can describe distortions from neutrality in addition to overall reductions in diversity.

By summarizing these effects in a time-dependent effective population size, our approach makes it possible to continue to use neutral methods for inference and estimation even in the presence of background selection, simply by allowing the population size to vary appropriately with time. For example, our results could be used in combination with a recent method developed by O'Fallon (2011) to incorporate a time-varying coalescent rate into genealogy samplers as a means to improve genealogical inference. This method considered a coalescence rate that declines linearly as time recedes into the past, but by instead incorporating the $N_e(t)$ we have calculated here, we can incorporate background selection and recombination into this and other existing neutral methods in a principled way.

We note that the $N_e(t)$ derived here is very similar to the purely nonrecombining case we analyzed in earlier work (Nicolaisen and Desai 2012). In both cases, $N_e(t)$ begins at the actual population size in the present, and then declines into the past before eventually reaching a long-term reduced size, $N_e(\infty)$. Although the explicit form of the $N_e(t)$ depends on the recombination rate, its qualitative features are similar to and lead to similar distortions to the asexual case. This similarity suggests that the general conclusions of earlier analyses of background selection in the nonrecombining case may often be qualitatively robust to the presence of recombination.

The reasons for this qualitative similarity (and the extent to which it holds) remains an open question. One appealing possibility is that local, closely linked genomic regions can be treated as effectively nonrecombining blocks, while loci separated by larger distances can be treated as freely recombining. Thus, a recombining population would be analogous to an asexual population with a particular "block length," which

would depend upon the strength of selection and the recombination rate. However, there are several problems with this intuition. In particular, it is not clear that there is a sufficiently sharp transition between regions that are effectively nonrecombining and those that are effectively freely recombining. Furthermore, the size of the effective block length may typically depend upon other parameters (such as the sample size).

We can naively attempt to quantify this similarity between a recombining population and an asexual population: If we define an effective genome size L^* (the "effectively nonrecombining block length", such that $U_d^* = \mu L^*$), and an effective selection strength s^*, then by demanding that the long-term effective population size $N_e(\infty)$ and the typical transition time to this long-term result are equivalent between the two populations, we arrive at

$$\begin{aligned} U_d^* &= U_d \left(1 - \frac{R/4}{s + R/4}\right) \\ s^* &= s \left(1 + \frac{R/4}{s + R/4}\right), \end{aligned}$$

Using these effective parameters, we find a close (though not identical) match between our results and the corresponding asexual model. However, it is not clear whether this rough equivalence or the effective parameters L^* and s^* have any predictive power beyond the statistics we have used here to define them. This remains an important topic for future work.

We note that our analysis rests on the assumption that mutation frequencies can be treated deterministically and that ancestral lineages can be treated independently. The implication of this latter assumption is that all pairs of lineages are considered independent and exchangeable, independent of the history of the sample. A consequence of this assumption is that, since all pairs of lineages are equally likely to coalesce, genealogies will be topologically neutral. As selection becomes weaker and these assumptions are violated, correlations between the lineages become important, topological distortions will arise, and our analysis breaks down. Furthermore, when this begins to occur, fluctuations in the sizes of lineages become very significant, and the deterministic assumption is also violated. Thus, methods capable of describing

these fluctuations are required to fully understand the effects of purifying selection and recombination in the weak selection regime. Very little is currently known about this regime, which is an important direction for future work.

We note that recent work has begun to address these fluctuations and the effects of weak purifying selection, but only in the purely asexual case. For example, (O'Fallon *et al.* 2010) developed a semi-analytical approach to understand these effects in the $Ns \approx 1$ regime. Their approach is to divide the population into a continuous distribution of fitness classes, calculate a corresponding ancestral fitness distribution and, in turn, coalescence rates. An alternative approach by Good *et al.* (2013) has suggested that the effects of many weakly selected mutations on sequence diversity are identical to the effects of fewer strongly selected mutations, making it possible to "map" weakly selected populations onto their equivalent strongly selected counterparts. An important question for future work is whether these weak selection methods may be extended to include recombination. A detailed understanding of the "effective" similarity between asexual and recombining populations, hinted at by our results, may potentially provide a way forward, by suggesting that these new asexual methods might also apply in the presence of recombination, with a suitable reinterpretation of the parameters.

5.4 Acknowledgements

We thank Benjamin Good for many useful discussions. This work was supported by the James S. McDonnell Foundation, the Harvard Milton Fund, and the Alfred P. Sloan Foundation. LEN is supported by the Department of Defense through the National Defense Science and Engineering Graduate Fellowship Program. Simulations were run on the Odyssey cluster supported by the FAS Sciences Division Research Computing Group at Harvard University.

SI A

Supplemental Information to Chapter Two

A.1 The Full Conditional Calculation

In the main text, we focused primarily on the non-conditional approximation to the coalescence probabilities, which led to our simple expression for the coalescence probabilities, Eq. 2.15. In this Supplementary Appendix, we show how this approximation can be relaxed in our lineage-structure framework by carrying out the full conditional calculation for some of the simplest possible cases. We use this to understand the structure of the conditional results and discuss the validity of the non-conditional approximation. We note that the full conditional result can also be obtained from the sum of ancestral paths approach by keeping the higher order terms in Eq. 2.56 of Appendix A, as described in Supplemental Information 1.4, and the validity of the non-conditional approximation can be directly assessed with that approach.

We begin by considering the full conditional result for the probability that two

---

From: **The Structure of Genealogies in the Presence of Purifying Selection: A "Fitness-Class Coalescent"** Aleksandra M. Walczak*, Lauren E. Nicolaisen*, Joshua B. Plotkin, and Michael M. Desai, *These authors contributed equally to this work

individuals both sampled from class k coalesce in class $k-2$. From Appendix A of the main text, we have

$$P_c^{k,k\to k-2} = I_x^{k-2} \int Q_{k,k}^{k-2}(t_1,t_2) \exp\left[-s(k-2)|t_1-t_2|\right] dt_1 dt_2. \tag{A.1}$$

In order to evaluate this integral, we need to determine the probability distribution of mutant timings $Q_{k,k}^{k-2}(t_1,t_2)$. The time t_1 is the sum of the time for one individual to have mutated from class $k-2$ to class $k-1$ plus the time for it to have mutated from class $k-1$ to class k, and analogously for t_2. However, in order for the two lineages to coalesce in class $k-2$, they must *not* have coalesced in class $k-1$. To illustrate the main point, we neglect the distortion in the mutant timings due to the fact that individuals did not coalesce in class k and focus only on the distortions due to the fact that coalescence did not occur in class $k-1$; if desired, the former distortion can also be included using analogous methods. We refer to the probability distribution of the times when these individuals mutated from class $k-1$ to class k conditional on them not having coalesced in class $k-1$ as $Q_{k,k}^{k-1}(t_1,t_2|nc)$. The distribution of the times for these individuals to then have mutated from class $k-2$ to class $k-1$ is then given by

$$Q_{1step}^{k-2}(t_1,t_2) = [s(k-1)]^2 e^{-s(k-1)(t_1+t_2)}. \tag{A.2}$$

Thus the distribution of t_1 and t_2 is given by

$$Q_{k,k}^{k-2}(t_1,t_2) = Q_{k,k}^{k-1}(t_1,t_2|nc) \star Q_{1step}^{k-2}(t_1,t_2), \tag{A.3}$$

where $\star$ indicates a convolution. Note that much of the time when the individuals did coalesce in class $k-1$, they did so because t_1 happened to be close to t_2 (since this increases the chance the two individuals mutated from the same lineage). Thus in $Q_{k,k}^{k-1}(t_1,t_2|nc)$, t_1 and t_2 are on average further apart than in $Q_{k,k}^{k-1}(t_1,t_2)$, and t_1 and t_2 are no longer independent random variables.

We now need to calculate $Q_{k,k}^{k-1}(t_1,t_2|nc)$. We have

$$Q_{k,k}^{k-1}(t_1,t_2|nc) = \frac{Q_{k,k}^{k-1}(t_1,t_2) - Q_{k,k}^{k-1}(t_1,t_2|c) P_c^{k,k\to k-1}}{1 - P_c^{k,k\to k-1}}, \tag{A.4}$$

where $Q_{k,k}^{k-1}(t_1,t_2|c)$ is the distribution of timings of mutations from class $k-1$ to k given that the lineages *do* coalesce in class $k-1$. Applying the general probability identity $P(t_1,t_2|c) = \frac{1}{P(c)}P(c|t_1,t_2)P(t_1,t_2)$, and reading off the coalescence probability given t_1 and t_2 from Eq. 2.13, we find that

$$Q_{k,k}^{k-1}(t_1,t_2|c) = \frac{I_x^{k-1}}{P_c^{k,k\to k-1}} Q_{k,k}^{k-1}(t_1,t_2) e^{-s(k-1)|t_1-t_2|}. \tag{A.5}$$

We therefore find

$$Q_{k,k}^{k-1}(t_1,t_2|nc) = \frac{1}{1-P_c^{k,k\to k-1}}\left[(sk)^2 e^{-sk(t_1+t_2)} - I_x^{k-1}(sk)^2 e^{-2k(t_1+t_2)} e^{-s(k-1)|t_1-t_2|}\right]. \tag{A.6}$$

Plugging this into our convolution formula for $Q_{k,k}^{k-2}(t_1,t_2)$ and evaluating the integrals by separating out the possible time orderings, we find

$$Q_{k,k}^{k-2}(t_1,t_2) = \frac{k^2[s(k-1)]^2}{1-P_c^{k,k\to k-1}} e^{-s(k-1)(t_1+t_2)}\left[\left(1-e^{-st_1}\right)\left(1-e^{-2t_2}\right) - \frac{I_x^{k-1}}{k-2}B\right], \tag{A.7}$$

where we have defined

$$\begin{aligned} B &= \frac{1}{(k-2)}\left[1 - e^{-2s\min(t_1,t_2)} - \frac{2}{k}\left(1-e^{-sk\min(t_1,t_2)}\right)\right. \\ &\quad \left. + \frac{1}{k}\left(1-e^{-2k|t_1-t_2|}\right)\left(e^{-2s\min(t_1,t_2)} - e^{-sk\min(t_1,t_2)}\right)\right]. \end{aligned} \tag{A.8}$$

We can now use this expression in Eq. SI 1.1 to calculate the coalescence probability $P_c^{k,k\to k-2}$. Since the result is tedious and does not further illuminate the structure of the full conditional calculation, we do not do so explicitly here, but the integrals are straightforward to evaluate with the methods we have used above.

To motivate the validity of the non-conditional approximation, we need to consider the full calculation going back one additional step. Thus we consider the probability that two individuals both sampled from class k coalesce in class $k-3$, $P_c^{k,k\to k-3}$. This will be given by

$$P_c^{k,k\to k-3} = \int Q_{k,k}^{k-3}(t_1,t_2)\frac{x^2}{h_{k-3}^2} f_{k-3}(x) e^{-s(k-3)|t_1-t_2|} dt_1 dt_2 dx, \tag{A.9}$$

where here $Q_{k,k}^{k-3}(t_1,t_2)$ is the distribution of the time at which the ancestors of the two sampled individuals originally mutated from class $k-3$ to class $k-2$, conditional on them not coalescing in classes $k-2$ or $k-1$.

We can calculate $Q_{k,k}^{k-3}(t_1,t_2)$ in the same way we calculated $Q_{k,k}^{k-2}(t_1,t_2)$. Explicitly,

$$Q_{k,k}^{k-3}(t_1,t_2) = Q_{k,k}^{k-2}(t_1,t_2|nc) \star Q_{1step}^{k-3}(t_1,t_2), \tag{A.10}$$

where analogously to the expression in the previous step

$$Q_{k,k}^{k-2}(t_1,t_2|nc) = \frac{1}{1-P_c^{k,k\to k-2}}\left[Q_{k,k}^{k-2}(t_1,t_2) - Q_{k,k}^{k-2}(t_1,t_2|c)P_c^{k,k\to k-2}\right]. \tag{A.11}$$

We note that $Q_{k,k}^{k-2}(t_1,t_2)$ is the expression in Eq. (A.7) we calculated above. As before, we have

$$Q_{k,k}^{k-2}(t_1,t_2|c)P_c^{k,k\to k-2} = I_x^{k-2}Q_{k,k}^{k-2}(t_1,t_2)e^{-s(k-2)|t_1-t_2|}, \tag{A.12}$$

hence we can write

$$Q_{k,k}^{k-2}(t_1,t_2|nc) = \frac{Q_{k,k}^{k-2}(t_1,t_2)}{1-P_c^{k,k\to k-2}}\left[1 - I_x^{k-2}e^{-s(k-2)|t_1-t_2|}\right]. \tag{A.13}$$

Plugging the above expression back into Eq. (A.10), we obtain

$$\begin{aligned} Q_{k,k}^{k-3}(t_1,t_2) &= \frac{s^2(k-1)^2k^2s^2(k-2)^2}{(1-P_c^{k,k\to k-1})(1-P_c^{k,k\to k-2})}e^{-s(k-2)(t_1+t_2)}\int_0^{t_2}\int_0^{t_1}e^{s(k-2)(y+z)}e^{s(k-1)(y+z)} \\ &\quad \times\left[1-I_x^{k-2}e^{-s(k-z)|y-z|}\right]\left[(1-e^{-sy})(1-e^{-sz}) - \frac{I_x^{k-1}}{k-2}B\right]. \end{aligned} \tag{A.14}$$

We could evaluate the integrals in the above expression for $Q_{k,k}^{k-3}(t_1,t_2)$ in the same way that we did in our calculation for $Q_{k,k}^{k-2}(t_1,t_2)$. We would then substitute this result for $Q_{k,k}^{k-3}(t_1,t_2)$ into an analogous calculation of $Q_{k,k}^{k-4}(t_1,t_2)$, and so on. In this way we can build up the full conditional results. The most useful way to go about this is to separate the results into powers of I_x, which is a small parameter related to the coalescent probability in each step. We see from the expression for $Q_{k,k}^{k-3}(t_1,t_2)$ that there is a term in $(I_x)^0$, which is exactly the non-conditional approximation. There are two terms involving $(I_x)^1$, and a single term involving $(I_x)^2$. In general, in the expression for $Q_{k,k}^{k-\ell}(t_1,t_2)$, we will have one $(I_x)^0$ term (which equals the result in the non-conditional approximation) plus ℓ terms proportional to I_x, $\binom{2}{\ell}$ terms proportional to $(I_x)^2$, and so on. Fortunately, the dependence on the population parameters is entirely contained within these powers of I_x. That is, the coefficients of these various powers of I_x depend *only* on k and ℓ, and not at all on the population

parameters N, s, and U_d. Thus we could simply calculate a table of coefficients once, and then would be able to understand all the distributions of mutant timings (and from this all the coalescent probabilities).

In practice, it is easier to make these full conditional calculations within the sum of ancestral paths approach. As we show in Supplemental Information 1.4, that approach leads naturally to a power series in I_x of exactly the form described above, in which the leading order term is the non-conditional approximation and the additional terms represent the conditional corrections. This calculation shows that provided $I_x \ll 1$, which is true provided our usual condition that $Nh_k sk \gg 1$ holds, these higher order terms are all small, and our non-conditional approximation is valid.

These full conditional results are, however, very complex and unilluminating. Therefore we focus here on understanding the general structure of these results, and on showing why the non-conditional approximation is good description of the distribution of mutation timings. We can see that at each step back through the fitness distribution, the probability distribution of times shifts from the non-conditional results by a factor which is roughly proportional to the coalescence probability at that step. That is, in general we have

$$Q_{k,k}^{k-\ell}(t_1,t_2) = \frac{1}{1-P_c^{k,k\to k-\ell}}\left[Q_{k,k}^{k-\ell}(t_1,t_2) - P_c^{k,k\to k-\ell}Q_{k,k}^{k-2}(t_1,t_2|c)\right]. \tag{A.15}$$

The first term in square brackets reflects the fact that the probability distribution at a given step conditional on non-coalescence at that step is almost equal to the unconditional probability distribution at that step. The second term represents the correction: note that it is proportional to the coalescence probability in that step, $P_c^{k,k\to k-\ell}$. The nature of the correction can be seen by plugging in the distribution of times conditional on coalescence, giving

$$Q_{k,k}^{k-\ell}(t_1,t_2) = \frac{Q_{k,k}^{k-\ell}(t_1,t_2)}{1-P_c^{k,k\to k-\ell}}\left[1 - I_x^{k-\ell}e^{-s(k-\ell)|t_1-t_2|}\right]. \tag{A.16}$$

We see that the correction acts to reduce the probability that $|t_1 - t_2|$ is small — that is, it makes it more likely that t_1 and t_2 are further apart, because this is more likely to be the case given that coalescence did not occur.

Since at each step the shift in the distribution of mutant timings is proportional to the coalescence probability, and the coalescence probability at each step is small, it seems clear that the non-conditional approximation where we simply ignore this shift in mutant timings is reasonable. However there is one potential caveat we must consider: although the shift in the distribution of mutation timings due to conditioning on non-coalescence is small *in each step*, we typically take many steps before the lineages coalesce. In fact, since the shift in mutation timings is proportional to the coalescence probability, and we typically go back a number of steps of order one over the coalescence probability, in principle the shifts in mutation timings could add up to a substantial shift.

Fortunately, there are three factors which prevent this from happening. First, the shift in mutation timings at each step is always to reduce the probability of times t_1 and t_2 where $|t_1 - t_2| \lesssim \frac{1}{(k-\ell)s}$. Since at each step ℓ is increasing, and the range of separations between mutation timings at which coalescence can happen is also increasing, the shifts in mutation timings from many steps ago are not a huge factor in determining coalescence probabilities in a particular step. That is, though the shifts in mutation timings add up over many steps, the shifts most relevant to the coalescent probability in a given step do not. Second, the coalescence probabilities at each step are different. This reduces the chance that we take enough steps to shift the overall mutation timings substantially by the time we coalesce. Finally, and most importantly, we will see that the there is a substantial probability that the ancestors of the two individuals sampled do not coalesce until they are in the most-fit class. This means that the total sum of coalescence probabilities (and hence the total possible weight in the shift of mutation timings) remains small even in the worst case where the two lineages do not coalesce for the maximum possible number of steps. The non-conditional approximation will always be good in the regime where this is true. All of these heuristic conclusions are reflected in the fact that the full conditional result we calculate in the sum of ancestral paths approach is equal to the non-conditional result plus corrections that are small provided $I_x \gg 1$.

A.2 The Non-conditional Distributions of Mutant Timings

Within the non-conditional approximation we need to calculate the distribution of mutant timings, as used in Eq. 2.48. Specifically, we need to calculate

$$Q_k^{k-\ell}(t) = Q_k^{k-1}(t) \star Q_{k-1}^{k-2}(t) \star Q_{k-2}^{k-3}(t) \star \ldots \star Q_{k-\ell+1}^{k-\ell}(t), \tag{A.17}$$

where $\star$ refers to a convolution and

$$Q_{k-\ell+1}^{k-\ell}(t) = s(k-\ell+1)e^{-s(k-\ell+1)t}, \tag{A.18}$$

as given by Eq. (2.6). In general, the convolution of n exponential distributions with parameters $\lambda_1 \ldots \lambda_n$ is given by

$$\sum_{i=0}^{n-1} \lambda_i e^{-\lambda_i t} \prod_{j=0,\neq i}^{n-1} \frac{\lambda_j}{\lambda_j - \lambda_i}. \tag{A.19}$$

Applying this identity with $\lambda_i = s(k-i)$, we find

$$Q_k^{k-\ell}(t) = \sum_{i=0}^{\ell-1} s e^{-s(k-i)t} \left(\frac{\prod_{j=0}^{\ell-1} k-j}{\prod_{j=0,\neq i}^{\ell-1} i-j} \right) \tag{A.20}$$

We can simplify this expression by noting that

$$\prod_{j=0}^{\ell-1}(k-j) = \frac{k!}{(k-\ell)!}, \tag{A.21}$$

and similarly that

$$\prod_{j=0,\neq i}^{\ell-1}(i-j) = i!(\ell-1-i)!(-1)^{\ell-1-i}. \tag{A.22}$$

This means we have

$$Q_k^{k-\ell}(t) = \sum_{i=0}^{\ell-1} s\ell e^{-s(k-i)t}(-1)^{\ell-i-1}\binom{\ell-1}{i}\binom{k}{k-\ell}. \tag{A.23}$$

We can evaluate this sum by recognizing the binomial expansion formula

$$(1+x)^n = \sum_{i=0}^{n} x^i \binom{n}{i}, \tag{A.24}$$

where we identify $x = -e^{st}$. We find

$$Q_k^{k-\ell}(t) = s\ell \binom{k}{\ell} e^{-skt} \left(e^{st}-1\right)^{\ell-1}. \tag{A.25}$$

More generally, we have

$$Q_a^b(t) = s(a-b) \binom{a}{b} e^{-sat} \left(e^{st}-1\right)^{a-b-1}. \tag{A.26}$$

A.3 General Coalescence Probabilities in the Non-conditional Approximation

The probability of coalescence for two individuals originally in two different classes k and k', as defined in Eq. 2.48 can be rewritten as

$$P_c^{k,k'\to k'-\ell} = \frac{1}{1+2Nh_{k-\ell}s(k-\ell)}\left[I_1+I_2\right], \tag{A.27}$$

where we have defined

$$I_1 = \int_0^\infty Q_{k'}^{k-\ell}(t_1)e^{-s(k-\ell)t_1}\int_0^{t_1} Q_k^{k-\ell}(t_2)e^{s(k-\ell)t_2}dt_2dt_1 \tag{A.28}$$

$$I_2 = \int_0^\infty Q_k^{k-\ell}(t_2)e^{-s(k-\ell)t_2}\int_0^{t_2} Q_{k'}^{k-\ell}(t_1)e^{s(k-\ell)t_1}dt_1dt_2. \tag{A.29}$$

Note that both I_1 and I_2 involve integrals of the form

$$I_a = \int_0^t Q_a^b(t')e^{sbt'}dt'. \tag{A.30}$$

Plugging in the results for the non-conditional distributions of mutant timings, Eq. SI 1.26, and making use of the binomial expansion formula for $(1+x)^n$ noted in Supplemental Information 1.2, we find this integral becomes

$$I_a = s(a-b)\binom{a}{b}\int_0^t e^{s(b-a)t'}\left(e^{st'}-1\right)^{a-b-1}dt' \tag{A.31}$$

$$= s(a-b)\binom{a}{b}\sum_{i=0}^{a-b-1}(-1)^{a-b-1+i}\binom{a-b-1}{i}\int_0^t e^{s(b-a+i)t'}dt' \tag{A.32}$$

$$= (a-b)\binom{a}{b}(-1)^{a-b}\sum_{i=0}^{a-b-1}\frac{(-1)^i}{a-b}\binom{a-b}{i}\left(e^{s(b-a+i)t}-1\right) \tag{A.33}$$

$$= \binom{a}{b}(-1)^{a-b}\sum_{i=0}^{a-b}(-1)^i\binom{a-b}{i}\left(e^{s(b-a+i)t}-1\right) \tag{A.34}$$

$$= \binom{a}{b}(-1)^{a-b}e^{s(b-a)t}\sum_{i=0}^{a-b}\left(-e^{st}\right)^i\binom{a-b}{i} \tag{A.35}$$

$$= \binom{a}{b}e^{s(b-a)t}\left(e^{st}-1\right)^{a-b}. \tag{A.36}$$

We now substitute this result for I_a into our expressions for I_1 and I_2. We note that both have terms of the form

$$I_b = \int_0^\infty Q_a^b(t)\binom{c}{b} e^{-sct}\left(e^{st}-1\right)^{c-b} dt. \tag{A.37}$$

Using similar manipulations to those above, we find

$$\begin{aligned} I_b &= (a-b)\tbinom{a}{b}\tbinom{c}{b}\textstyle\int_0^\infty e^{-s(a+c)t}\left(e^{st}-1\right)^{a+c-2b-1}dt && \text{(A.38)}\\ &= s(a-b)\tbinom{a}{b}\tbinom{c}{b}(-1)^{a+c-1}\textstyle\sum_{i=0}^{a+c-2b-1}\tbinom{a+c-2b-1}{i}(-1)^i\int_0^\infty e^{-s(a+c-i)t}dt && \text{(A.39)}\\ &= (a-b)\tbinom{a}{b}\tbinom{c}{b}(-1)^{a+c-1}\textstyle\sum_{i=0}^{a+c-2b-1}(-1)^i\tbinom{a+c-2b-1}{i}\frac{1}{a+c-i}. && \text{(A.40)} \end{aligned}$$

Using the partial fraction decomposition

$$\frac{1}{\binom{n+x}{n}} = \sum_{i=1}^{n}(-1)^{i-1}\binom{n}{i}\frac{i}{x+i}, \tag{A.41}$$

we find

$$I_b = \frac{\frac{a-b}{a+c-2b}\binom{a}{b}\binom{c}{b}(-1)^{a+c}}{\binom{-2b-1}{a+c-2b}} = \frac{\frac{a-b}{a+c-2b}\binom{a}{b}\binom{c}{b}(-1)^{2b}}{\binom{a+c}{a+c-2b}}. \tag{A.42}$$

We can now use this result for I_b to determine I_1 and I_2, and hence compute $P_c^{k,k'\to k'-\ell}$. We find

$$P_c^{k,k'\to k'-\ell} = \frac{1}{1+2Nh_{k-\ell}s(k-\ell)}\frac{\binom{k'}{k-\ell}\binom{k}{k-\ell}}{\binom{k+k'}{2\ell+k'-k}}. \tag{A.43}$$

As we noted in the main text, this is just

$$P_c^{k,k'\to k-\ell} = \frac{1}{1+2Nh_{k-\ell}s(k-\ell)}A_\ell^{k,k'}, \tag{A.44}$$

with $A_\ell^{k,k'}$ as defined in Eq. 2.16. Note that when $k=k'$, this result simplifies to $P_c^{k,k\to k-\ell}$ as defined in the main text, as expected.

A.4 Computing Sums of Ancestral Paths

In this appendix, we describe the calculation of $\phi_k^{k'}(\ell)$ using the sum of ancestral paths approach.

A.4.1 Calculation of $\phi_k^k(3)$

We begin by considering a simpler specific case, where $k = k'$ and $\ell = 3$. There are a total of $\binom{6}{3} = 20$ possible ancestral paths by which two individuals sampled from class k can coalesce in class $k-3$. These can be separated into four types, according to whether the two ancestral lineages were ever together in classes $k-1$ or $k-2$. We can list all paths of each type, using the notation that A is a mutation event in the first lineage, and B is a mutation event in the second lineage. We have

$$\underbrace{\begin{pmatrix} ABABAB \\ ABABBA \\ ABBAAB \\ ABBABA \\ BAABAB \\ BAABBA \\ BABAAB \\ BABABA \end{pmatrix}}_{\binom{2}{1}\binom{2}{1}\binom{2}{1}=8 \text{ ways}} \quad \underbrace{\begin{pmatrix} ABAABB \\ ABBBAA \\ BAAABB \\ BABBAA \end{pmatrix}}_{\binom{2}{1}\left(\binom{4}{2}-\binom{2}{1}\binom{2}{1}\right)=4 \text{ ways}} \quad \underbrace{\begin{pmatrix} AABBAB \\ AABBBA \\ BBAAAB \\ BBAABA \end{pmatrix}}_{\binom{2}{1}\left(\binom{4}{2}-\binom{2}{1}\binom{2}{1}\right)=4 \text{ ways}} \quad \underbrace{\begin{pmatrix} AAABBB \\ AABABB \\ BBBAAA \\ BBABAA \end{pmatrix}}_{\binom{6}{3}-\text{others}=4\text{ways}}.$$

The probabilities of all paths of a particular type are identical. We can calculate the probability of each of the four types of paths using the same logic as outlined in the main text. We find

$$P(AAABBBc) = I_x^{k-3}\frac{k(k-1)(k-2)}{8(2k-1)(2k-3)(2k-5)}\left(1-I_x^k\right), \tag{A.45}$$

$$P(AABBABc) = I_x^{k-3}\frac{k(k-1)(k-2)}{8(2k-1)(2k-3)(2k-5)}\left(1-I_x^k\right)\left(1-I_x^{k-1}\right), \tag{A.46}$$

$$P(ABAABBc) = I_x^{k-3}\frac{k(k-1)(k-2)}{8(2k-1)(2k-3)(2k-5)}\left(1-I_x^k\right)\left(1-I_x^{k-2}\right), \tag{A.47}$$

$$P(ABABABc) = I_x^{k-3}\frac{k(k-1)(k-2)}{8(2k-1)(2k-3)(2k-5)}\left(1-I_x^k\right)\left(1-I_x^{k-1}\right)\left(1-I_x^{k-2}\right). \tag{A.48}$$

Summing over all the possible paths, we find

$$\phi_k^k(3)=I_{k-3}\frac{\binom{k}{k-3}\binom{k}{k-3}}{\binom{2k}{6}}\left[1-\frac{\binom{2}{1}\binom{4}{2}}{\binom{6}{3}}I_{k-1}-\frac{\binom{2}{1}\binom{4}{2}}{\binom{6}{3}}I_{k-2}+\frac{\binom{2}{1}\binom{2}{1}\binom{2}{1}}{\binom{6}{3}}I_{k-1}I_{k-2}\right]. \tag{A.49}$$

We now pause to consider the form of the probabilities of each type of ancestral path. These probabilities differ only by factors of $(1-I_x^{k-i})$. One such factor arises each time the two ancestral lineages are together in class $k-i$. In other words, we can rewrite the probability of each path as the probability of an undistorted path (defined to be a path in which the contributions due to the possibility of coalescence in previous classes are neglected), times a correction for each class in which the two lineages are together:

$$\begin{aligned}
P(AAABBBc) &= P(\text{Undistorted Path})\big(1-I_x^k\big) && \text{(A.50)}\\
P(AABBABc) &= P(\text{Undistorted Path})\big(1-I_x^k\big)\big(1-I_x^{k-1}\big) && \text{(A.51)}\\
P(ABAABBc) &= P(\text{Undistorted Path})\big(1-I_x^k\big)\big(1-I_x^{k-2}\big) && \text{(A.52)}\\
P(ABABABc) &= P(\text{Undistorted Path})\big(1-I_x^k\big)\big(1-I_x^{k-1}\big)\big(1-I_x^{k-2}\big). && \text{(A.53)}
\end{aligned}$$

By definition, the "undistorted path" probability is the probability neglecting the contributions due to the possibility of coalescence in previous steps, and is therefore the same for all paths. We have

$$\begin{aligned}
P(\text{Undistorted Path}) &= \frac{k(k-1)(k-2)k(k-1)(k-2)}{2k(2k-1)(2k-2)(2k-3)(2k-4)(2k-5)}I_x^{k-\ell} && \text{(A.54)}\\
&= \frac{\frac{k!}{(k-3)!}\frac{k!}{(k-3)!}}{\frac{2k!}{(2k-6)!}}I_x^{k-\ell}. && \text{(A.55)}
\end{aligned}$$

Using these results, we can write $\phi_k^k(3)$ as

$$\begin{aligned}
\phi_k^k(3) = {} & [\#\text{ of Paths}]P(\text{Undistorted Path})\big[F_k(1-I_x^k)+F_{k,k-1}(1-I_x^k)(1-I_x^{k-1})\\
& +F_{k,k-2}(1-I_x^k)(1-I_x^{k-2})+F_{k,k-1,k-2}(1-I_x^k)(1-I_x^{k-1})(1-I_x^{k-2})\big], \qquad \text{(A.56)}
\end{aligned}$$

where we have defined $F_{\{a\}}$ to be the fraction of paths that are together in the set of classes $\{a\}$ (and are not together in any other class).

A.4.2 Calculation of $\phi_{k'}^k(\ell)$

We now use this approach to calculate the coalescence probability in the general case. The probability of any particular ancestral path from k and k' to $k-\ell$ is the product

of the individual probabilities of each mutational step that makes up this path. Each such individual probability consists of three parts: a numerator, which depends only on the current class of the lineage that mutates, divided by a denominator, which depends only on the sum of the current set of classes for both lineages, times a correction factor of $(1 - I_x^{k-i})$ if the two lineages are in the same class at that step.

Although in each ancestral path the mutations will occur in a different order, all paths will ultimately consist of the same set of mutations ($k' \to k'-1 \to \ldots \to k-\ell$ and $k \to k-1 \to \ldots \to k-\ell$). Therefore, regardless of the path taken, the product of the numerators from each step will be identical. Similarly, the sum of the current set of classes will begin at $k'+k$, and decrement by one each time a deleterious mutation occurs, until both lineages are in the final class ($k'+k \to k'+k-1 \to \ldots \to 2k-2\ell$). Therefore, regardless of the path taken, the product of the denominators from each step will also be identical. Therefore, the paths will differ only by the correction factor $(1-I_x^{k-i})$ for each class in which the two ancestral lineages are together. This means that, analogous to the case of $\phi_k^k(3)$ we described above, the probability of each path is the probability of an "undistorted path" times the appropriate correction factor. The probability of the undistorted path is

$$P(\text{Undistorted Path}) = \tfrac{k'(k'-1)\ldots(k-\ell+1)k(k-1)\ldots(k-\ell+1)}{(k'+k)(k'+k-1)\ldots(2k-2\ell+1)} I_x^{k-\ell}. \tag{A.57}$$

We can now sum up all possible paths to obtain

$$\begin{aligned} \phi_{k'}^k(\ell) \;=\; & [\#\text{ of Paths}]P(\text{Undistorted Path})\big[F_\emptyset + \textstyle\sum_{i=0}^{\ell} F_{k-i}(1-I_x^{k-i}) \\ & + \textstyle\sum_{i=0}^{\ell-1}\sum_{j>i}^{\ell} F_{k-i,k-j}(1-I_x^{k-i})(1-I_x^{k-j}) \\ & + \textstyle\sum_{i=0}^{\ell-2}\sum_{j>i}^{\ell-1}\sum_{m>j}^{\ell} F_{k-i,k-j,k-m}(1-I_x^{k-i})(1-I_x^{k-j})(1-I_x^{k-m}) + \ldots\big], \end{aligned} \tag{A.58}$$

where as before $F_{\{a\}}$ is the fraction of paths that are together in the set of classes $\{a\}$ (and are not together in any other class). Note that there are a total of $\ell+1$ terms in this equation, representing the possibility that the two lineages can be together in anywhere from 0 to ℓ of the classes. We can rearrange these terms to write

$$\begin{aligned} \phi_{k'}^k(\ell) \;=\; & [\#\text{ of Paths}]P(\text{Undistorted Path})\big[1 - \textstyle\sum_{i=0}^{\ell} G_{k-i} I_x^{k-i} \\ & + \textstyle\sum_{i=0}^{\ell-1}\sum_{j>i}^{\ell} G_{k-i,k-j} I_x^{k-i} I_x^{k-j} \\ & - \textstyle\sum_{i=0}^{\ell-2}\sum_{j>i}^{\ell-1}\sum_{m>j}^{\ell} G_{k-i,k-j,k-m} I_x^{k-i} I_x^{k-j} I_x^{k-m} + \ldots\big], \end{aligned} \tag{A.59}$$

where we have defined $G_{\{a\}}$ to be the fraction of paths that are together in *at least* the set of classes $\{a\}$.

We can evaluate each of these factors of G. For example, the fraction of paths that are together in class $k-i$ equals the number of ways for the two lineages to descend from classes k' and k to be together in class $k-i$, $\binom{k'-k+2i}{i}$, times the number of ways for the two lineages to descend from class $k-i$ to be together in class $k-\ell$, $\binom{2i-2\ell}{i-\ell}$, divided by the total number of ways for the two lineages to descend from classes k' and k to be together in $k-\ell$, $\binom{k'-k+2\ell}{\ell}$. Using this logic, we find

$$\begin{aligned}\phi^k_{k'}(\ell) &= [\#\text{ of Paths}]P(\text{Undistorted Path}) \\ &\times\left[1-\textstyle\sum_{i=0}^{\ell-1}\frac{\binom{k'-k+2i}{i}\binom{2\ell-2i}{\ell-i}}{\binom{k'-k+2\ell}{\ell}}I_x^{k-i}+\sum_{i=0}^{\ell-2}\sum_{j>i}^{\ell-1}\frac{\binom{k'-k+2i}{i}\binom{2j-2i}{j-i}\binom{2\ell-2j}{\ell-j}}{\binom{k'-k+2\ell}{\ell}}I_x^{k-i}I_x^{k-j}\ldots\right].\end{aligned} \tag{A.60}$$

The total number of paths is $\binom{k'-k+2\ell}{\ell}$, so we finally find that the full probability of coalescence in class $k-\ell$ is

$$\begin{aligned}\phi^{k'}_k(\ell) &= I_x^{k-\ell}\frac{\binom{k'}{k-\ell}\binom{k}{k-\ell}}{\binom{k'+k}{k'-k+2\ell}}\left[1-\sum_{i=0}^{\ell-1}\frac{\binom{k'-k+2i}{i}\binom{2\ell-2i}{\ell-i}}{\binom{k'-k+2\ell}{\ell}}I_x^{k-i}+\right. \\ &\left.\sum_{i=0}^{\ell-2}\sum_{j>i}^{\ell-1}\frac{\binom{k'-k+2i}{i}\binom{2j-2i}{j-i}\binom{2\ell-2j}{\ell-j}}{\binom{k'-k+2\ell}{\ell}}I_x^{k-i}I_x^{k-j}-\ldots\right].\end{aligned} \tag{A.61}$$

This is Eq. 2.56 from the main text. Note that it equals our non-conditional result for $P_c^{k,k'\to\ell}$ times a correction factor. There are a total of $\ell+1$ terms in this correction factor. This full correction factor can be arbitrarily complex for large ℓ, so we do not write out a general form here. However, it is straightforward to calculate for any values of k, k', and ℓ; a Mathematica script to do so is available on request.

A.5 The Correspondence between Steptimes and Real Times

In this Supplementary Appendix, we calculate the correspondence between steptimes and the actual times measured in generations. Our goal is to calculate the probability distribution of real coalescence times, $\psi(t|k,k',\ell)$, given that individuals were initially in classes k and k' and coalesced in class $k-\ell$.

To begin, we neglect the coalescence time within class $k-\ell$, and consider the distribution of the time at which an ancestor of one of the two sampled individuals first mutated from class $k-\ell$ to class $k-\ell+1$. We refer to this as $\psi_1(t|k,k',\ell)$. We first calculate the joint distribution of the times at which both ancestors mutated out of the class, $R^{k-\ell}_{k,k'}(t_1,t_2)$. Conditional on coalescence in class $k-\ell$, $R^{k-\ell}_{k,k'}(t_1,t_2)$, is given by the probability of t_1 and t_2 and coalescence divided by the total probability of coalescence. That is,

$$R(t_1,t_2) = \frac{P(coal|t_1,t_2)P(t_1,t_2)}{P(coal)}. \tag{A.62}$$

Substituting in the relevant expressions from the main text, this gives

$$R^{k-\ell}_{k,k'}(t_1,t_2) = \frac{1}{A^{k,k'}_{\ell}} Q^{k-\ell}_{k,k'}(t_1,t_2) e^{-s(k-\ell)|t_1-t_2|}. \tag{A.63}$$

The time at which the first ancestor mutated out of class $k-\ell$ is the longer of the two times t_1 and t_2,

$$\psi(t|k,k',\ell) = \left[\int_0^t R^{k-\ell}_{k,k'}(t_1,t)dt_1 + \int_0^t R^{k-\ell}_{k,k'}(t,t_2)dt_2\right]. \tag{A.64}$$

Substituting in our expression for $R^{k-\ell}_{k,k'}(t_1,t_2)$ and carrying out the integrals as in Supplemental Information 1.3, we find

$$\psi_1(t|k,k',\ell) = s\pi_d e^{-s(k'+k)t}(e^{st}-1)^{\pi_d-1}\binom{k'+k}{\pi_d}, \tag{A.65}$$

where we have used $\pi_d = k'-k+2\ell$.

We can alternatively calculate $\psi_1(t|k,k',\ell)$ using our sum of ancestral paths approach. As before, we imagine two individuals sampled from classes k and k' and condition on them coalescing in class $k-\ell$. Consider a case where $k \neq k'$. Then

the first event in the history of these two individuals must be a deleterious mutation. Since these mutations happen at rate sk and sk' in each lineage, the distribution of times since this mutation occurred in one of the two ancestral lineages is

$$P(t) = s(k + k')e^{-s(k+k')t}. \tag{A.66}$$

With probability $\frac{k'}{k+k'}$, this mutation is in the lineage sampled from class k', in which case the two lineages are now in classes k and $k' - 1$. Alternatively, the mutaion occurred in the lineage sampled from k and the lineages are in classes $k - 1$ and k'.

We can now consider the time to the next event backwards in time. If the two lineages are in the same class (but not yet in class $k - \ell$), the distribution of times to the next deleterious mutation event is somewhat shorter, because we are conditioning on coalescence not occuring. However, provided that $2sk_1 \gg \frac{1}{Nh_k}$ (the condition we are already making elsewhere), this shortening of the time will be a small correction and neglecting it is a good approximation.

Making this approximation, the rate at which the next deleterious mutation event occurs when the two lineages are in classes k_1 and k_2 is just $s(k_1 + k_2)$. Regardless of the order in which these mutations happen between the two lineages, this sum is simply decreased by s at each step. This will continue until the both ancestral lineages are in class $k - \ell$. Therefore, the distribution of times until the original mutation out of class $k - \ell$ is given by:

$$\psi_1(t|k',k,\ell)=s(k'+k)e^{-s(k'+k)t}\star s(k'+k-1)e^{-s(k'+k-1)t}\star\ldots\star s(2k-2\ell+1)e^{-s(2k-2\ell+1)t}. \tag{A.67}$$

This can be written as

$$\psi_1(t|k', k, \ell) = \lambda_0 e^{-\lambda_0 t} \star \lambda_1 e^{-\lambda_1 t} \star \ldots \star \lambda_{k'-k+2\ell-1} e^{-\lambda_{k'-k+2\ell-1} t}, \tag{A.68}$$

where we have defined:

$$\lambda_i = s(k' + k - i). \tag{A.69}$$

We can compute this convolution as in Supplemental Information 1.2 (compare to Eq. SI 1.17 for $Q^{2k-2\ell}_{k+k'}(t)$). We find

$$\psi_1(t|k, k', \ell) = s\pi_d e^{-s(k'+k)t}(e^{st} - 1)^{\pi_d - 1}\binom{k' + k}{\pi_d}, \tag{A.70}$$

identical to the result of our lineage structure calculation above.

A.5.1 Distribution of Coalescence Times

To calculate the correspondence between steptimes and real times, we now need to add the time it takes two individuals two coalesce in class $k-\ell$, which we refer to as $\psi_2(t|k,k',\ell)$, to the time it took them both to get to that class, $\psi_1(t|k,k',k-\ell)$. The rate of coalescence once in class $k-\ell$ is $\frac{1}{Nh_{k-\ell}}$, so we have

$$\psi_2(t|k',k,\ell) = (2s(k-\ell) + 1/Nh_{k-\ell})\, e^{-[2s(k-\ell)+1/Nh_{k-l}]t}. \tag{A.71}$$

Putting this together, the full distribution of times since coalescence is

$$\psi(t|k',k,\ell) = \psi_1(t|k',k,\ell) \star \psi_2(t|k',k,\ell). \tag{A.72}$$

Carrying out this convolution (and expanding the binomial factor $(e^{st}-1)^{\pi_d-1}$ in ψ_1), we find

$$\psi(t|k',k,\ell)=\sum_{i=0}^{\pi_d-1} s\pi_d(-1)^{\pi_d-i-1}\binom{\pi_d-1}{i}\binom{k'+k}{\pi_d}\frac{B}{A-B}\left(e^{-sBt}-e^{-sAt}\right), \tag{A.73}$$

where we have defined $A \equiv k'+k-i$ and $B \equiv 2\,(k-\ell) + \frac{1}{Nsh_{k-\ell}}$.

A.6 An Alternative Approach to Neutral Diversity

Instead of calculating the distribution of neutral heterozygosity by first computing the distribution of real times, we could alternatively incorporate neutral mutations directly into the sum of ancestral paths framework. This completely bypasses the correspondence with real coalescence times. To do this, we characterize ancestral paths not only by the ordering of deleterious mutation and coalescence events, but also by the ordering of neutral mutations. This means that if we sample two individuals A and B, there are five types of events that can happen in their ancestral paths: a deleterious mutation (DM) in A or in B, a neutral mutation (NM) in either A or in B, and or a coalescence (C) event (if A and B are currently in the same class).

We now imagine that we sample two individuals from classes k and k', and that they coalesce in class $k-\ell$. Our goal is to calculate the probability distribution of π_n given k, k', and ℓ, $\rho(\pi_n|k,k',\ell)$. We will find it helpful to divide the five types of events that can occur into two classes: neutral mutations on the one hand, and deleterious mutations or coalescence (which we call "steps") on the other. We begin by computing the probability that a given number of NMs occur before the next DM or C events (i.e. the number of neutral mutations that occur at this "step"). We have

$$P(\text{a NMs, then DM in } k' \text{ or } k'|k',k) = \left(\frac{\frac{2U_n}{s}}{k'+k+\frac{2U_n}{s}}\right)^a \frac{k+k'}{k'+k+\frac{2U_n}{s}}, \qquad \text{(A.74)}$$

where we have made our usual assumption that $Nh_k sk \gg 1$, allowing us to neglect the rates of coalescence events (when $k=k'$) in writing this expression.

This probability only depends on the sum of the current classes the individulas are in. At each subsequent step, regardless of the path taken, this sum of the classes will decrease by one. Therefore, the probability that a_i neutral mutations occur at step i is independent of the path taken. This observation allows us to calculate the probability that a given total number of neutral mutations have occurred since coalescence. We first calculate the probability that a given number of neutral mutations have occurred since the first deleterious mutation out of the $k-\ell$ class. We will add in the additional neutral mutations once in the $k-\ell$ class at the end.

In order for π_n neutral mutations to have occurred since the first deleterious mutation out of class $k-\ell$, we require that a_0 mutations occurred at the first step, a_1 mutations occurred at the second step, and so on, such that $a_0+a_1+\ldots+a_{k'-k+2\ell-1} = \pi_n$. This gives

$$\rho(\pi_n{=}X|k',k,\ell){=}\frac{\frac{(k'+k)!}{(2k-2\ell)!}}{\frac{(\frac{2U_n}{s}+k'+k)!}{(\frac{2U_n}{s}+2k-2\ell)!}}\sum_{|\vec{a}|=X}\left(\tfrac{2U_n/s}{2U_n/s+k+k'}\right)^{a_0}\ldots\left(\tfrac{2U_n/s}{2U_n/s+2k-2l+1}\right)^{a_{k'-k+2l-1}}. \tag{A.75}$$

We can define $x \equiv 2U_n/s + k + k'$, recognize $\pi_d = k' - k + 2\ell$, and relabel the a_i as

$$a_0{\to}X{-}b_0, \quad a_1{\to}b_0{-}b_1, \quad \ldots \quad a_{\pi_d-2}{\to}b_{\pi_d-3}{-}b_{\pi_d-2}, \quad a_{\pi_d-1}{\to}b_{\pi_d-2}. \tag{A.76}$$

This gives

$$\begin{aligned}\rho(\Pi_n{=}X|k',k,\ell) \quad &= \quad \frac{\binom{k'+k}{\pi_d}}{\binom{\frac{2U_n}{s}+k'+k}{\pi_d}}\left(\tfrac{2U_n}{s}\right)^X\left(\tfrac{1}{x}\right)^X\textstyle\sum_{b_0=0}^{X}\left(\frac{x}{x-1}\right)^{b_0} \\ &\quad \textstyle\sum_{b_1=0}^{b_0}\left(\frac{x-1}{x-2}\right)^{b_1}\cdots\sum_{b_{\pi_d-2}=0}^{b_{\pi_d-3}}\left(\frac{x-\pi_d+2}{x-\pi_d+1}\right)^{b_{\pi_d-2}}.\end{aligned} \tag{A.77}$$

To simnplify this expression, it is helpful to define a function f such that:

$$\begin{aligned}\mathbf{f}(A,B) \quad &\equiv \quad \left(\tfrac{1}{x}\right)^X\textstyle\sum_{b_0=0}^{X}\left(\frac{x}{x-1}\right)^{b_0} \\ &\quad \textstyle\sum_{b_1=0}^{b_0}\left(\frac{x-1}{x-2}\right)^{b_1}\cdots\sum_{b_{A-1}=0}^{X}\left(\frac{x-A+1}{x-A}\right)^{b_0}\sum_{b_A=0}^{b_{A-1}}\left(\frac{x-A}{x-B}\right)^{b_A}\end{aligned} \tag{A.78}$$

In other words, $\mathbf{f}\,(A,B)$ is a set of A nested sums, each of the same form, except for the final sum, which can have a different denominator. Using this definition, we have

$$P(\Pi_n = X|k',k,\ell) = \frac{\binom{k'+k}{\pi_d}}{\binom{\frac{2U_n}{s}+k'+k}{\pi_d}}\left(\frac{2U_n}{s}\right)^X\mathbf{f}\,(\pi_d-2,\pi_d-1)\,. \tag{A.79}$$

The virtue of this definition is that this sum can be solved recursively. We have

$$\sum_{b_A=0}^{b_{A-1}}\left(\frac{x-A}{x-B}\right)^{b_A} = \frac{x-B}{A-B} - \frac{x-A}{A-B}\left(\frac{x-A}{x-B}\right)^{b_{A-1}}. \tag{A.80}$$

Therefore we have

$$\mathbf{f}\,(A,B) = \frac{x-A}{B-A}\mathbf{f}\,(A-1,B) - \frac{x-B}{B-A}\mathbf{f}\,(A-1,A)\,. \tag{A.81}$$

Repeatedly inserting this result yields:

$$\begin{aligned}
\mathbf{f}\left(A,A+1\right) &\rightarrow \tfrac{(x-A)(x-A-1)}{1}\left(\tfrac{\mathbf{f}(A-1,A+1)}{x-A-1}-\tfrac{\mathbf{f}(A-1,A)}{x-A}\right)\\
\mathbf{f}\left(A,A+1\right) &\rightarrow \tfrac{(x-A+1)(x-A)(x-A-1)}{2}\left[\tfrac{\mathbf{f}(A-2,A+1)}{x-A-1}-\tfrac{2\mathbf{f}(A-2,A)}{x-A}+\tfrac{\mathbf{f}(A-2,A-1)}{x-A+1}\right]\\
&\vdots\\
\mathbf{f}\left(A,A+1\right) &\rightarrow (m+1)\tbinom{x-A-1+m}{m+1}\textstyle\sum_{i=0}^{m}\tfrac{(-1)^{i+m}}{x-A-1+i}\tbinom{m}{i}\mathbf{f}(A-m,A+1-i).
\end{aligned} \tag{A.82}$$

Note that $\mathbf{f}(-1,B)=1/B^X$, since there are no more sums to compute. Thus, for $m=A+1$ we have

$$\mathbf{f}\left(A,A+1\right)=(A+2)\binom{x}{A+2}\sum_{i=0}^{A+1}\frac{(-1)^{i+A+1}}{(x-A-1+i)^{X+1}}\binom{A+1}{i}. \tag{A.83}$$

Relabeling the sum and taking $A=\pi_d-2$, we have

$$\mathbf{f}\left(\pi_d-2,\pi_d-1\right)=\pi_d\binom{x}{\pi_d}\sum_{i=0}^{\pi_d-1}\frac{(-1)^{i}}{(x-i)^{X+1}}\binom{\pi_d-1}{i}. \tag{A.84}$$

We can now substitute these results into our expression for π_n, to find

$$\rho_1(\Pi_n{=}X|k',k,\ell){=}\pi_d\tbinom{k'+k}{\pi_d}\left(\tfrac{2U_n}{s}\right)^X\textstyle\sum_{i=0}^{\pi_d-1}\tfrac{(-1)^i}{(2U_n/s+k+k'-i)^{X+1}}\tbinom{\pi_d-1}{i} \tag{A.85}$$

Note, however, that this is only the distribution of neutral mutations since the first deleterious mutation out of class $k-l$. It is also possible for neutral mutations to occur prior to the coalescence event. Adding in this factor, we find

$$\begin{aligned}
\rho(\Pi_n{=}X|k',k,\ell) \;=\; & \pi_d\tbinom{k'+k}{\pi_d}\textstyle\sum_{i=0}^{\pi_d-1}(-1)^i\tbinom{\pi_d-1}{i}\\
& \times\textstyle\sum_{X=0}^{\pi_n}\tfrac{(2U_n/s)^X}{(2U_n/s+k+k'-i)^{X+1}}\left(\tfrac{2N_{k-l}U_n}{1+2N_{k-l}U_n+2N_{k-l}s(k-l)}\right)^{\pi_n-X}.
\end{aligned} \tag{A.86}$$

Rearranging this expression gives

$$\rho(\pi_n|k',k,\ell){=}\textstyle\sum_{i=0}^{\pi_d-1}\pi_d(-1)^{\pi_d-i-1}\tbinom{\pi_d-1}{i}\tbinom{k'+k}{\pi_d}\tfrac{B}{A-B}\left(\tfrac{(\frac{2U_n}{s})^{\pi_n}}{(\frac{2U_n}{s}+B)^{\pi_n+1}}-\tfrac{(\frac{2U_n}{s})^{\pi_n}}{(\frac{2U_n}{s}+A)^{\pi_n+1}}\right), \tag{A.87}$$

where we have defind

$$A=k'+k-i,\qquad B=2\left(k-\ell\right)+\frac{1}{Nsh_{k-l}}, \tag{A.88}$$

identical to our earlier result.

SI B

Supplemental Information to Chapter Five

B.1 Approximations

In our derivation of the time-dependent effective population size, we have made two key approximations. First, we have assumed that lineages and allele frequencies may be treated as effectively deterministic. Second, we have assumed that the ancestral fitness distributions at different sites may be treated as independent. These two approximations are prevalent in the history of background selection, and form the basis for many of the strong-selection results currently in use (Charlesworth 2012; Charlesworth *et al.* 1993; Hudson and Kaplan 1995b). In this Supplemental Information, we discuss these two approximations in detail.

B.1.1 The Deterministic Approximation

One of the central assumptions of background selection is that the population may be treated as approximately deterministic. This implies that frequencies may be assumed

From: **Distortions in Genealogies due to Purifying Selection and Recombination** Lauren E. Nicolaisen and Michael M. Desai, *Genetics* **195**, 1 (2013).

to be at mutation-selection balance, and that lineages may be described using deterministic equations such as that used to derive Eq. (5.1). In general, this assumption will hold when the strength of selection is sufficiently strong that it dominates the effects of drift (or analogously, when lineages are selected against sufficiently strongly that they never grow to a substantial fraction of the population). As a result, we expect the deterministic approximation to hold roughly when $Nse^{-U_d/(s+R/2)} \gg 1$. This approximation forms the foundation for previous results in background selection, including the structured coalescent results of Zeng and Charlesworth (2011) and the original background selection formulae from Charlesworth *et al.* (1993) and Hudson and Kaplan (1995b).

The main difference between the classic background selection analysis and our analysis is that we include the transient period during which deleterious alleles may segregate in the population prior to being removed by selection. The traditional analysis assumes that this time-period is sufficiently small relative to the total coalescence time that it can be neglected. In general, the time-scale of this transition is roughly of order $1/s$, and therefore, by definition, should be small relative to the typical coalescence times, $\approx N_e$, whenever the deterministic approximation holds.

However, in practice, as seen in Figures 5.3-5, the deterministic approximation is still reasonable even when the time-scale of the transition begins to represent a significant fraction of the total coalescence times. Thus, by incorporating this transition time, we are able to more accurately describe the distribution of coalescence times and other statistics. This allows us to capture the distortions that begin to arise as a consequence of this transition period, and thus to qualitatively understand how selection distorts the shapes of genealogies, and how this depends upon the parameters involved. Even when this effect is small, by taking advantage of the fact that, in the presence of recombination, sites far away from one another become effectively independent, it may be possible to detect even small differences with enough sequence data. We note, however, that our method is only able to account for the distortions that arise due to this transition period, and not the additional effects that arise from fluctuations. As $N_e s$ becomes smaller, our analysis begins to break down as fluctuations in the population become very strong. When this happens, additional

distortions (including topological distortions) arise which we are not able to capture with our analysis.

The breakdown of the deterministic approximation when $N_e s \approx 1$ has been discussed in several notable studies considering the weak selection regime (Barton and Etheridge 2004; O'Fallon *et al.* 2010). Earlier studies have suggested that the deterministic approximation is reasonable for the calculation of pairwise coalescence times when $N_e s > 3$ (Barton and Etheridge 2004; Charlesworth 2012), which is consistent with our findings in Figure 5.4. However, it is unclear whether such a precise threshold would remain accurate for more extreme parameter combinations, where additional logarithmic corrections could arise.

B.1.2 The Independent-Sites Approximation

The second key approximation made in the main text is that we may treat the ancestral fitness distribution at each site as independent. In other words, we assume that the joint ancestral fitness distribution across all sites is equal to the product of the ancestral fitness distribution at each site, $P_{k_1,k_2,k_3\ldots k_L}(t) = P_{k_1}(t)P_{k_2}(t)\ldots P_{k_L}(t)$, where k_i is either 0 or 1, indicating whether a mutation exists at site i.

In an asexual population, this holds whenever the deterministic approximation is valid. However, in the presence of recombination, correlations will exist between neighboring sites. This is a consequence of the fact that, when an ancestral recombination event occurs between the focal site and multiple index sites, all of those sites will now be randomly chosen from the population at the same time, and thus will all be 'reset' to the steady state mutation-selection balance simultaneously. Thus, sites that share the same history will be correlated.

However, this effect will be small provided that the deterministic approximation is valid ($N_e s \gg 1$) and that the probability of a mutation at any given site is small ($\mu/s \ll 1$). This approximation is prominent in previous literature on background selection, and is discussed in detail in the appendix of Hudson and Kaplan (1995b). In order to justify this approximation, we will show that, provided the conditions stated above hold, the joint fitness distribution at two loci are approximately in-

dependent, i.e. $P_{k_1,k_2}(t) = P_{k_1}(t)P_{k_2}(t) + \mathcal{O}(\frac{\mu^2}{s^2})$. The same argument can then be extended to additional loci.

We denote the ancestral fitness distribution of an individual as $P_{ij}(t)$, where i and j represent whether a mutation exists at two sites of distances x_1 and x_2 from the focal site, respectively. We know from the main text that, to first order in μ/s:

$$\begin{aligned} P_{00}(t) + P_{01}(t) &= 1 - \frac{\mu}{s}\left(\frac{rx_1}{rx_1+s} + \frac{s}{rx_1+s}e^{-st-rx_1t}\right) \\ P_{00}(t) + P_{10}(t) &= 1-\frac{\mu}{s}\left(\frac{rx_2}{rx_2+s}+\frac{s}{rx_2+s}e^{-st-rx_2t}\right). \end{aligned}$$

We can now write out the backwards-in-time master equation for $P_{00}(t)$, again keeping only first-order terms in μ, s, rx_1, and rx_2:

$$\begin{aligned} P_{00}(t+1) = {} & P_{00}(t)(1-rx_1(1-f_{00})-r(x_2-x_1)(1-f_{00}-f_{10})) \\ & +P_{01}(t)\tfrac{f_{00}}{f_{01}}(\mu + rx_1f_{01} + r(x_2-x_1)(f_{01}+f_{11})) \\ & +P_{10}(t)\tfrac{f_{00}}{f_{10}}(\mu+rx_1f_{10}) \\ & +P_{11}(t)\tfrac{f_{00}}{f_{11}}(rx_1f_{11}). \end{aligned}$$

Making the continuous approximation this becomes:

$$\begin{aligned} \tfrac{dP_{00}(t)}{dt} &= -rx_2P_{00}(t) + rx_1f_{00}(P_{00}(t)+P_{01}(t)+P_{10}(t)+P_{11}(t)) \\ &\quad +r(x_2-x_1)\Big(P_{00}(t)(f_{00}+f_{10}) + \tfrac{f_{00}}{f_{01}}P_{01}(t)(f_{01}+f_{11})\Big) + \mu f_{00}\left(\tfrac{P_{01}(t)}{f_{01}} + \tfrac{P_{10}(t)}{f_{10}}\right) \\ &= -(rx_2+2s-2\mu)P_{00}(t)+rx_1\left(1-\tfrac{\mu}{s}\right)^2 \\ &\quad +r(x_2-x_1)\left(1-\tfrac{\mu}{s}\right)\left(1-\tfrac{\mu}{s}\left(\tfrac{rx_1}{rx_1+s}+\tfrac{s}{rx_1+s}e^{-st-rx_1t}\right)\right) \\ &\quad +s\left(1-\tfrac{\mu}{s}\right)\left(2-\tfrac{\mu}{s}\left(\tfrac{rx_1}{rx_1+s}+\tfrac{s}{rx_1+s}e^{-st-rx_1t}\right)-\tfrac{\mu}{s}\left(\tfrac{rx_2}{rx_2+s}+\tfrac{s}{rx_2+s}e^{-st-rx_2t}\right)\right). \end{aligned}$$

Solving this to first order in μ/s:

$$\begin{aligned} P_{00}(t) &= 1 - \frac{\mu}{s}\left(\frac{rx_1}{rx_1+s} + \frac{s}{rx_1+s}e^{-st-rx_1t}\right) - \frac{\mu}{s}\left(\frac{rx_2}{rx_2+s} + \frac{s}{rx_2+s}e^{-st-rx_2t}\right) + \mathcal{O}\left(\frac{\mu^2}{s^2}\right) \\ P_{01}(t) &= \tfrac{\mu}{s}\left(\tfrac{rx_2}{rx_2+s}+\tfrac{s}{rx_2+s}e^{-st-rx_2t}\right)+\mathcal{O}\left(\tfrac{\mu^2}{s^2}\right) \\ P_{10}(t) &= \tfrac{\mu}{s}\left(\tfrac{rx_1}{rx_1+s}+\tfrac{s}{rx_1+s}e^{-st-rx_1t}\right)+\mathcal{O}\left(\tfrac{\mu^2}{s^2}\right) \\ P_{11}(t) &= o\left(\tfrac{\mu^2}{s^2}\right). \end{aligned}$$

Thus, we see that $P_{ij}(t) = P_i(t)P_j(t)+\mathcal{O}\left(\mu^2/s^2\right)$, such that the sites are approximately independent. We note, however, that this independence does not hold to higher-order

in $\frac{\mu}{s}$, and corrections would be required to accurately capture the joint ancestral probability at those orders. Thus, the independence approximation will only strictly hold when $\mu/s \ll 1$, and when the deterministic approximation holds.

We note that this approximation is discussed in detail in the appendix of HUDSON and KAPLAN (1995b). They provide an analogous derivation of the joint mutation probability at two loci (see Equation A10), and similarly find that sites may be treated as independent provided the deterministic approximation holds and $\mu/s \ll 1$.

B.2 Incorporating Back Mutations

In our derivation of the time-dependent effective population size, we have neglected the effect of back mutations. In practice, back mutations only introduce terms of higher-order in μ/s, and thus are of negligible contribution in the regime we consider. However, it is straightforward to incorporate these terms into our analysis, which we do here.

First, we consider the steady-state distribution of mutations at a single site. This is determined by the solution to the equations:

$$\begin{aligned} f_1 &= \frac{f_1(1-s)}{\overline{\omega}}(1-\mu_b) + \frac{f_0}{\overline{\omega}}\mu_f \\ f_0 &= \frac{f_1(1-s)}{\overline{\omega}}\mu_b + \frac{f_0}{\overline{\omega}}(1-\mu_f), \end{aligned}$$

where μ_f and μ_b are the forward and back mutation rates, respectively. This yields:

$$\begin{aligned} f_1 &= \frac{s+\mu_b(1-s)+\mu_f - \sqrt{(s+\mu_b(1-s)+\mu_f)^2 - 4s\mu_f}}{2s} \\ f_0 &= \frac{s-\mu_b(1-s)-\mu_f + \sqrt{(s+\mu_b(1-s)+\mu_f)^2 - 4s\mu_f}}{2s}. \end{aligned}$$

When $\mu_b = 0$, these reduce to the usual mutation-selection balance results, $f_1 = \mu_f/s$ and $f_0 = 1 - \mu_f/s$. Furthermore, if we define $\mu_f \equiv \mu$ and $\mu_b \equiv c\mu$, and expand this result in orders of μ/s, we see that:

$$\begin{aligned} f_1 &= \frac{\mu}{s} - \frac{\mu^2}{s^2}c(1-s) + \frac{\mu^3}{s^3}(c^2(1-s)^2 - c(1-s))\ldots \\ f_0 &= 1 - \frac{\mu}{s} + \frac{\mu^2}{s^2}c(1-s) - \frac{\mu^3}{s^3}(c^2(1-s)^2 - c(1-s))\ldots \end{aligned}$$

Thus, we see that incorporating back mutations leads to a correction of order μ^2/s^2. As a consequence, the effect of back mutations is negligible in the regime we consider. However, we may derive Equation 5.2 from the main text including them. We have that:

$$\frac{dP_{mut}(t)}{dt} = -\left(rx + \frac{\mu_f N f_0}{N f_1} + \frac{\mu_b N f_1}{N f_0}\right) P_{mut}(t) + rx f_1 + \frac{\mu_b N f_1}{N f_0}$$

Solving this yields:

$$P_{mut}(x,t) = \frac{rxf_1 + \frac{\mu_b f_1}{f_0}}{rx + \frac{\mu_f f_0}{f_1} + \frac{\mu_b f_1}{f_0}} + \frac{\mu_f f_0 - \mu_b f_1}{rx + \frac{\mu_f f_0}{f_1} + \frac{\mu_b f_1}{f_0}} e^{-\left(rx + \frac{\mu_f f_0}{f_1} + \frac{\mu_b f_1}{f_0}\right)t}.$$

which replaces Equation 5.2 in the main text. Similarly, Equation 5.1 may be recovered by substituting $rx \to r(x_i, x_f)$. We note that these equations are identical to those given in the main text to leading-order in μ/s, and thus back mutations represent only a small correction to our results in the regime we consider.

Bibliography

Abramowitz, M., and I. A. Stegun, 1965 *Handbook of Mathematical Functions*. Dover, New York.

Aquadro, C. F., S. F. Desse, M. M. Bland, C. H. Langley, and C. C. Laurie-Ahlberg, 1986 Molecular population genetics of the alcohol dehydrogenase gene region of drosophila melanogaster. Genetics **114**: 1165–1190.

Austerlitz, F., B. Jung-Muller, B. Godelle, and P.-H. Gouyon, 1997 Evolution of coalescence times, genetic diversity and structure during colonization. Theoretical Population Biology **51**: 148–164.

Barton, N. H., and A. M. Etheridge, 2004 The effect of selection on genealogies. Genetics **166**: 1115–1131.

Betancourt, A. J., J. J. Welch, and B. Charlesworth, 2009 Reduced effectiveness of selection caused by a lack of recombination. Current Biology **19**: 655–660.

Burridge, C., D. Craw, D. Fletcher, and J. Waters, 2008 Geological dates and molecular rates: fish dna sheds light on time dependency. Molecular Biology and Evolution **25**: 624.

Charlesworth, B., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. Genetical Research **63**: 213–227.

Charlesworth, B., 2010 Molecular population genomics: a short history. Genetics research **92**: 397–411.

Charlesworth, B., 2012 The effects of deleterious mutations on evolution at linked sites. Genetics **190**: 5–22.

Charlesworth, B., 2013 Background selection 20 years on: The Wilhelmine E. Key 2012 invitational lecture. Journal of Heredity .

Charlesworth, B., and D. Charlesworth, 1997 Rapid fixation of deleterious alleles can be caused by muller's ratchet. Genetical Research **70**: 63–73.

Charlesworth, B., M. Morgan, and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134**: 1289.

Charlesworth, B., *et al.*, 1996 Background selection and patterns of genetic diversity in drosophila melanogaster. Genetical Research **68**: 131–150.

Charlesworth, D., B. Charlesworth, and M. T. Morgan, 1995 The pattern of neutral molecular variation under the background selection model. Genetics **141**: 1619–1632.

Comeron, J. M., and M. Kreitman, 2002 Population, evolutionary and genomic consequences of interference selection. Genetics **161**: 389–410.

Comeron, J. M., A. Williford, and R. M. Kliman, 2008 The hill-robertson effect: Evolutionary consequences of weak selection and linkage in finite populations. Heredity **100**: 19–31.

Coop, G., and R. C. Griffiths, 2004 Ancestral inference on gene trees under selection. Theroetical Population Biology **66**: 219–232.

Desai, M. M., and D. S. Fisher, 2007 Beneficial mutation-selection balance and the effect of linkage on positive selection. Genetics **176**: 1759–1798.

Desai, M. M., L. E. Nicolaisen, A. M. Walczak, and J. B. Plotkin, 2012 The structure of allelic diversity in the presence of purifying selection. Theor. Pop. Biol. **81**: 144–157.

Eriksson, A., B. Mehlig, M. Rafajlovic, and S. Sagitov, 2010 The total branch length of sample genealogies in populations of variable size. Genetics **186**: 601–611.

Etheridge, A. M., and R. C. Griffiths, 2009 A coalescent dual process in a moran model with genic selection. Theoretical Population Biology **75**: 320–330.

Etheridge, A. M., R. C. Griffiths, and J. E. Taylor, 2010 A coalescent dual process in a moran model with genic selection, and the lambda coalescent limit. Theoretical Population Biology **78**: 77–92.

Ethier, S. N., and T. G. Kurtz, 1987 The infinitely-many alleles model with selection as a measure-valued diffusion. Stochastic Models in Biology, Lecture Notes in Biomathematics **70**: 72–86.

Ethier, S. N., and T. G. Kurtz, 1994 Convergence to fleming-viot processes in the weak atomic topology. Stochastic Processes and their Applications **54**: 1–27.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theoretical Population Biology **3**: 87–112.

Ewens, W. J., 2004 *Mathematical Population Genetics: I. Theoretical Introduction*. Springer, New York, NY.

Ewens, W. J., and W.-H. Li, 1980 Frequency spectra of neutral and deleterious alleles in a finite population. Journal of Mathematical Biology **10**: 155–166.

Eyre-Walker, A., and P. Keightley, 1999 High genomic deleterious mutation rates in hominids. Nature **397**: 344–347.

Fay, J., G. Wyckoff, and C. Wu, 2001 Positive and negative selection on the human genome. Genetics **158**: 1227.

Fisher, R., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Fu, Y., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking, and background selection. Genetics **147**: 915–925.

Gessler, D. D. G., 1995 The constraints of finite size in asexual populations and the rate of the ratchet. Genetical Research **66**: 241–253.

Good, B. H., A. M. Walczak, R. A. Neher, and M. M. Desai, 2013 Interference limits resolution of selection pressures from linked neutral diversity. arXiv preprint arXiv:1306.1215 .

Gordo, I., and B. Charlesworth, 2000a The degeneration of asexual haploid populations and the speed of muller's ratchet. Genetics **154**: 1379–1387.

Gordo, I., and B. Charlesworth, 2000b On the speed of muller's ratchet. Genetics **156**: 2137–2140.

Gordo, I., A. Navarro, and B. Charlesworth, 2002 Muller's ratchet and the pattern of variation at a neutral locus. Genetics **161**: 835–848.

Griffiths, R. C., 1983 Allele frequencies with genic selection. Journal of Mathematical Biology **17**: 1–10.

Grote, M. N., and T. P. Speed, 2002 Approximate ewens formulae for symmetric overdominance selection. Annals of Applied Probability **12**: 637–663.

Haddrill, P., L. Loewe, and B. Charlesworth, 2010 Estimating the parameters of selection on nonsynonymous mutations in drosophila pseudoobscura and d. miranda. Genetics **185**: 1381–1396.

Hahn, M. W., 2008 Toward a selection theory of molecular evolution. Evolution **62**: 255–265.

Haigh, J., 1978 The accumulation of deleterious genes in a population-muller's ratchet. Theor Pop Biol **14**: 251–267.

Haldane, J., 1919 The combination of linkage values and the calculation of distances between the loci of linked factors. J. Genet **8**: 309–320.

Haldane, J. B. S., 1927 A mathematical theory of natural and artificial selection, part v: selection and mutation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 23. Cambridge Univ Press, 838–844.

Hartl, D. L., 1988 *A Primer of Population Genetics*. Sinauer Associates, Sunderland, MA.

Hermisson, J., O. Redner, H. Wagner, and E. Baake, 2002 Mutation-selection balance: ancestry, load, and maximum principle. Theor Pop Biol **62**: 9–46.

Hill, W., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genetical Research **8**: 269–294.

Ho, S., M. Phillips, A. Cooper, and A. Drummond, 2005 Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Molecular biology and evolution **22**: 1561.

Ho, S. Y., R. Lanfear, L. Bromham, M. J. Phillips, J. Soubrier, *et al.*, 2011 Time-dependent rates of molecular evolution. Mol Ecol **20**.

Hudson, R., and N. Kaplan, 1994 Gene trees with background selection. Non-neutral evolution: theories and molecular data : 140–153.

Hudson, R., and N. Kaplan, 1995a The coalescent process and background selection. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **349**: 19–23.

Hudson, R., and N. Kaplan, 1995b Deleterious background selection with recombination. Genetics **141**: 1605–1617.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theoretical population biology **23**: 183–201.

Hudson, R. R., *et al.*, 1990 Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology **7**: 44.

Joyce, P., 1995 Robustness of the ewens sampling formula. Journal of Applied Probability **32**: 609–622.

Joyce, P., and S. Tavare, 1995 The distribution of rare alleles. Journal of Mathematical Biology **33**: 602–618.

Kaiser, V. B., and B. Charlesworth, 2009 The effects of deleterious mutations on evolution in non-recombining genomes. Trends in Genetics **25**: 9–12.

Kaplan, N. L., T. Darden, and R. R. Hudson, 1988 The coalescent process in models with selection. Genetics **120**: 819–829.

Karlin, S., and J. McGregor, 1972 Addendum to a paper of w. ewens. Theoretical Population Biology **3**: 113–116.

Keightley, P., and A. Eyre-Walker, 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics **177**: 2251–2261.

Kim, Y., and W. Stephan, 2002 Recent applications of diffusion theory to population genetics. In M. Slatkin and M. Veuille, editors, *Modern Developments in Theoretical Population Genetics: The Legacy of Gustave Malecot*. Oxford University Press, Oxford, UK.

Kimura, M., 1955 Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbor Symposia on Quantitative Biology **20**: 33–53.

Kimura, M., and T. Maruyama, 1966 The mutational load with epistatic gene interactions in fitness. Genetics **54**: 1337.

Kingman, J. F. C., 1982 The coalescent. Stochastic Processes and their Applications **13**: 235–248.

Kreitman, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of drosophila melanogaster. Nature **304**: 412–417.

Krone, S. M., and C. Neuhauser, 1997 Ancestral processes with selection. Theoretical Population Biology **51**: 210–237.

Kuhner, M. K., 2009 Coalescent genealogy samplers: windows into population history. Trends in Ecology & Evolution **24**: 86–93.

Li, W.-H., 1977 Maintenance of genetic variability under mutation and selection pressures in a finite population. PNAS **74**: 2509–2513.

Li, W.-H., 1978 Maintenance of genetic variability under the joint effect of mutation, selection and random drift. Genetics **90**: 349–382.

Li, W.-H., 1979 Maintenance of genetic variability under the pressure of neutral and deleterious mutations in a finite population. Genetics **92**: 647–667.

Loewe, L., and B. Charlesworth, 2007 Background selection in single genes may explain patterns of codon bias. Genetics **175**: 1381–1393.

Lohmueller, K., A. Albrechtsen, Y. Li, S. Kim, T. Korneliussen, *et al.*, 2011 Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS genetics **7**: e1002326.

Lohse, K., R. J. Harrison, and N. H. Barton, 2011 A general method for calculating likelihoods under the coalescent process. Genetics **189**: 977–987.

Malécot, G., 1941 Etude mathématique des populations mendéliennes. Annales de l'Université de Lyon, Sciences (A-4) : 45–60.

McVean, G. A. T., and B. Charlesworth, 2000 The effects of hill-robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155**: 929–944.

McVicker, G., D. Gordon, C. Davis, and P. Green, 2009 Widespread genomic signatures of natural selection in hominid evolution. PLoS genetics **5**: e1000471.

Neuhauser, C., and S. M. Krone, 1997 The genealogy of samples in models with selection. Genetics **145**: 519–534.

Nicolaisen, L., and M. Desai, 2012 Distortions in genealogies due to purifying selection. Molecular Biology and Evolution **29**: 3589–3600.

Nielsen, R., and D. M. Weinreich, 1999 The age of nonsynonymous and synonymous mutations in animal mtdna and implications for the mildly deleterious theory. Genetics **153**: 497–506.

Nordborg, M., 2001 *Coalescent theory*. Wiley Online Library.

Nordborg, M., B. Charlesworth, and D. Charlesworth, 1996 The effect of recombination on background selection. Genetical Research **67**: 159–174.

O'Fallon, B., 2011 A method for accurate inference of population size from serially sampled genealogies distorted by selection. Molecular biology and evolution **28**: 3171–3181.

O'Fallon, B., J. Seger, and F. Adler, 2010 A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. Mol. Biol. Evol. **27**: 1162.

O'Fallon, B. D., 2010 A method to correct for the effects of purifying selection on genealogical inference. Mol. Biol. Evol. **27**: 2406.

Penny, D., 2005 Relativity for molecular clocks. Nature **436**: 183.

Polanski, A., and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics **165**: 427–436.

Pool, J. E., I. Hellmann, J. D. Jensen, and R. Nielsen, 2010 Population genetic inference from genomic sequence variation. Genome research **20**: 291–300.

Powell, J. R., 1994 Molecular techniques in population genetics: a brief history. In *Molecular ecology and evolution: Approaches and applications*. Springer, 131–156.

Przeworski, M., B. Charlesworth, and J. Wall, 1999 Genealogies and weak purifying selection. Molecular Biology and Evolution **16**: 246–252.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132**: 1161–1176.

Seger, J., W. A. Smith, J. J. Perry, J. Hunn, Z. A. Kaliszewska, *et al.*, 2010 Gene genealogies strongly distorted by weakly interfering mutations in constant environments. Genetics **184**: 529–545.

Stephens, J. C., and M. Nei, 1985 Phylogenetic analysis of polymorphic dna sequences at the adh locus indrosophila melanogaster and its sibling species. Journal of molecular evolution **22**: 289–300.

Stephens, M., 2008 Inference under the coalescent. Handbook of Statistical Genetics, Third Edition : 878–908.

Tajima, F., 1983 Evolutionary relationship of dna sequences in finite populations. Genetics **105**: 437–460.

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

Tavaré, S., 2004 Part i: Ancestral inference in population genetics. In *Lectures on probability theory and statistics*. Springer, 1–188.

Wakeley, J., 2009 *Coalescent Theory, an Introduction*. Roberts and Company, Greenwood Village, CO.

Wakeley, J., 2010 Natural selection and coalescent theory. Evolution since Darwin: the first 150 years **150**: 119–149.

Wakeley, J., and J. Hey, 1997 Estimating ancestral population parameters. Genetics **145**: 847.

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

Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. Theoretical population biology **7**: 256–276.

Watterson, G. A., 1978 The homozygosity test of neutrality. Genetics **88**: 405–417.

Weir, J., and D. Schluter, 2008 Calibrating the avian molecular clock. Molecular Ecology **17**: 2321–2328.

Williamson, S., and M. Orive, 2002 The genealogy of a sequence subject to purifying selection at multiple sites. Mol. Biol. Evol. **19**: 1376.

Woodhams, M., 2006 Can deleterious mutations explain the time dependency of molecular rate estimates? Mol. Biol. Evol. **23**: 2271.

Wright, S., 1931 Evolution in mendelian populations. Genetics **16**: 97.

Zemlak, T. S., E. M. Habit, S. J. Walde, C. Carrea, and D. E. Ruzzante, 2010 Surviving historical patagonian landscapes and climate: molecular insights from galaxias maculatus. BMC Evol Biol **10**: 67.

Zeng, K., 2012 A coalescent model of background selection with recombination, demography and variation in selection coefficients. Heredity **110**: 363–371.

Zeng, K., and B. Charlesworth, 2011 The joint effects of background selection and genetic recombination on local gene genealogies. Genetics **189**: 251–266.