A Continuous-State Coalescent and the Impact of Weak Selection on the Structure of Gene Genealogies

Brendan D. O'Fallon,*[,1] Jon Seger,[2] and Frederick R. Adler[2]

[1]Department of Genome Sciences, University of Washington
[2]Department of Biology, University of Utah

***Corresponding author:** E-mail: brendano@u.washington.edu.

Associate editor: Rasmus Nielsen

Abstract

Coalescent theory provides an elegant and powerful method for understanding the shape of gene genealogies and resulting patterns of genetic diversity. However, the coalescent does not naturally accommodate the effects of heritable variation in fitness. Although some methods are available for studying the effects of strong selection ($Ns \gg 1$), few tools beyond forward simulation are available for quantifying the impact of weak selection at many sites. Here, we introduce a continuous-state coalescent capable of accurately describing the distortions to genealogies caused by moderate to weak natural selection affecting many linked sites. We calculate approximately the full distribution of pairwise coalescent times, the lengths of coalescent intervals, and the time to the most recent common ancestor of a sample. Weak selection ($Ns \approx 1$) is found to substantially decrease the tree depth, primarily through a shortening of the lengths of the basal coalescent intervals. Additionally, we demonstrate that only two parameters, population size and the variance of the distribution describing fitness heritability, are sufficient to describe most changes.

Key words: coalescent, weak selection, gene genealogies, interference, multisite selection.

Research article

Introduction

Understanding the manner in which natural selection affects patterns of genetic variation is of fundamental importance in population genetics. Coalescent theory, first described by Kingman (1982a,b,c), provides an elegant framework for this endeavor by describing the shapes of gene genealogies and resulting patterns of variation. However, because the coalescent relies on the assumption that individuals in the population do not differ in their expected reproductive success, the theory does not easily accommodate natural selection. Although many authors have extended coalescent theory to include various forms of selection, most have focused on selective schemes with only a small number of segregating alleles, often two (Krone and Neuhauser 1997; Neuhauser and Krone 1997; Barton and Etheridge 2004; Coop and Griffiths 2004; Wakeley 2008). Although two-allele models aid our understanding in situations such as the selective sweep of an advantageous allele, many populations, particularly those with large sizes or high mutation rates, contain loci that simultaneously segregate many alleles. The coalescent process is less well understood in these situations. In this paper, we demonstrate that relatively weak natural selection affecting multiple linked sites can significantly distort the shapes of gene genealogies from the predictions of neutral and two-allele models, and we develop methods that accurately predict these distortions.

Previous work on the coalescent with selection has focused on the limiting cases of strong and weak selection. If selection is strong, then allele frequencies either remain constant or change deterministically over time. The population may be thought of as several subpopulations, each corresponding to an allele (or more accurately, a fitness variant). Within subpopulations, expected reproductive success is equal, and thus, coalescent theory requires only modest modification to accurately describe these cases (Kaplan et al. 1988; Wakeley 2008). This approach may be applied to a variety of selective schemes (such as overdominance and balancing selection). However, natural selection must be fairly strong for this approximation to hold, particularly if many alleles or loci are considered (Barton and Navarro 2002; Navarro and Barton 2002).

If natural selection is weaker, then allele frequencies fluctuate over time, and it is necessary to incorporate these fluctuations to accurately model genealogies. Barton and Etheridge (2004) and Barton et al. (2004), extending a model first put forth by Kaplan et al. (1988), used a diffusion approximation to describe the probability that an allelic class had a particular frequency at some time in the past and then described relationships between genes conditional on the allelic frequencies. The method worked well for the one-locus, two-allele case but was numerically difficult to extend to cases involving more loci or alleles. Hudson and Kaplan (1994, 1995) analyzed a model that tracked a larger number of mutational classes that could be applied to weak selection. Selection was presumed to act in a multiplicative manner based on the number of mutations experienced by a particular sequence, and frequencies of these mutational classes were assumed to be Poisson distributed and constant. However, for weaker selection coefficients, the distribution of allelic classes will not remain constant, and the Poisson approximation becomes inaccurate.

At loci that harbor more than a few alleles, the combined effect of many mutations may distort genealogical structure even if each mutation has only a small fitness effect (Przeworski et al. 1999; McVean and Charlesworth 2000;

Williamson and Orive 2002; Maia et al. 2004; Comeron et al. 2008). Quantitative analysis is difficult in this regime because both selection and drift are important and fitness variants may arise and be lost frequently. Results have primarily been obtained through simulation studies, both forward simulation (Golding 1997; McVean and Charlesworth 2000; Williamson and Orive 2002; Maia et al. 2004) and simulated reconstruction of the genealogy itself, using the ancestral selection graph (ASG; Krone and Neuhauser 1997; Neuhauser and Krone 1997; Przeworski et al. 1999). These studies have primarily concluded that weak selection has only a modest effect on genealogical structure and one that is maximized for intermediate levels of selection. However, forward simulation techniques cannot handle realistic population sizes, particularly when the entire genealogy must be tracked. Additionally, the ASG becomes inaccurate if multiple sites combine to yield large selection coefficients, thus limiting the total strength of selection that can be modeled (see Przeworski et al. 1999; some recent modifications to the ASG allow for stronger selection, e.g., Slade 2000).

At their core, many of the studies above involve the "structured coalescent" (Nordborg 1997) in which the population is divided into a number of discrete groups, usually representing allelic or fitness states. Within a given allelic, class individuals are identical, and thus, the neutral coalescent accurately describes the history of samples within groups, whereas the mutational regime describes movement between groups. In the case of weak selection, the number of potential states grows rapidly, and tracking the size of groups and movement of lineages among groups approaches intractability. To address these concerns, we have developed a model that assumes an infinite number of fitness states. In this case, the matrix describing transitions between groups becomes a continuous function (similar to a dispersal kernel), and the probability that a lineage is in a certain state some number of generations ago is given by a continuous probability density function. Despite this difference, the basic approach remains unchanged. We track the probability that a lineage is in a certain state some number of generations ago and then calculate the probability that two individuals share a parent in the previous generation by integrating over the distribution of potential states. If the population size is much larger than the sample size, then this pairwise coalescent rate is sufficient to describe the ancestry of the entire sample.

In this paper, we utilize the continuous approximation to examine the impact of weak selection operating at multiple sites on the structure of a genealogy. Our model tracks only the expected reproductive success of individuals, thus fitness is a quantitative trait and individuals are endowed not with genotypes or allelic states but with a single (nonnegative) real number describing expected reproductive success. We investigate how the distribution of ancestral fitnesses changes as one looks deeper into the past, and how this influences the probability that two randomly selected individuals first shared a common ancestor at a certain time. These calculations are used to find the distribution of pairwise coalescence times, the distribution of lengths of coalescent intervals, and the time to the most recent common ancestor (TMRCA) of a sample of genes. The results are also compared with simulations using a more realistic finite-sites model of fitness variation. In addition, we demonstrate that a single parameter describing the variance in fitness heritability in a single generation is sufficient to describe most distortions to the genealogies brought about by selection.

Methods

Model Description

We begin by describing a simple population genetic model where individuals are endowed with a genome consisting of a finite number of sites. We then demonstrate how a model that tracks only the relative fitness of individuals can be used to approximate the discrete-sites model. Using the simpler relative fitness model, we address genealogical structure in three steps. First, we calculate the probability that a randomly chosen lineage (the series of ancestors of an individual chosen from the "present" generation) has fitness w at a given generation in the past. Second, we calculate the probability that two lineages, with fitnesses drawn from the probability distribution calculated in the first step, first shared a common parent t generations ago. Finally, we use the calculations to derive the expected lengths of coalescent intervals and TMRCA for a sample of arbitrary size. We verify our assumptions through comparison to forward simulations, both of our continuous model and a more realistic model with discrete number of sites. Unless otherwise noted, we use "fitness" to mean relative fitness or, equivalently, an individual's expected number of offspring.

Consider an asexual population of constant size N with nonoverlapping generations. Each individual contains a nonrecombining genome of L sites where each site may exist in one of two possible states. Each site is mutated independently with probability μ each generation. An individual's absolute fitness is determined by the total number of sites that differ from a predetermined most-fit genotype. Specifically, if the genome differs at n sites, absolute fitness is given by e^{-sn}. Each new generation is populated by selecting individuals in proportion to their absolute fitness, and if selected, a parent produces a single offspring. The parental generation is sampled repeatedly and with replacement until exactly N offspring exist. Because sites may mutate to both more- and less fit states, the model encompasses both beneficial and deleterious mutations. Similar models have been analyzed by a number of authors, including McVean and Charlesworth (2000), Comeron and Kreitman (2002) Rouzine et al. (2003), and Seger et al. (2010).

The primary approximation we use in this work is to track only the evolution of the relative fitnesses of individuals. Specifically, we assume that there is some function $f(w_o;w_p,\tau^2)$ that describes the probability that an individual has relative fitness w_o conditional on its parent having fitness w_p. The parameter τ^2 describes the variance of this distribution; if $\tau^2 = 0$, then offspring have fitness identical to their parents and the model collapses to neutrality. If f is

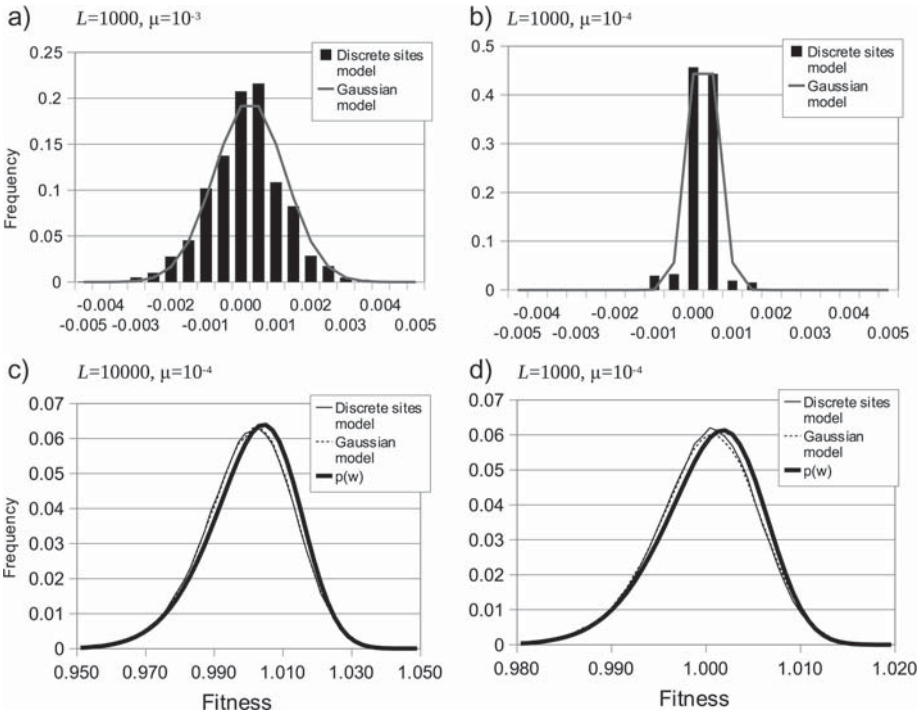

FIG. 1. (*a*, *b*) Distributions of the difference of offspring and parent relative fitness (the function *f*) simulated in the discrete-sites model (black bars) and the Gaussian approximation used in the Gaussian model. (*c*, *d*) Simulated stationary distribution of relative fitness values for the discrete-sites and Gaussian models and the function $p(w)$ used to approximate it in the numerical methods. $N = 1,000$, $s = 10^{-3}$ in all cases.

known, then offspring fitnesses may easily be generated by drawing a single random variable from *f* instead of simulating mutation at many independent sites. In the Appendix, we derive the mean and the variance of *f* for the discrete-sites model. To first order in *s* and μ, $\tau^2 = L\mu s^2$ and is independent of parental fitness. Although we have been unable to derive a closed form for *f*, it is approximately Gaussian if somewhat leptokurtic for $L\mu < 1$ (fig. 1*a* and *b*). In the calculations and simulations below, we use a Gaussian function for *f*, and we refer to this model as the "Gaussian model." Although the true *f* for the discrete-sites model is not exactly Gaussian (only a finite number of fitnesses are possible), we demonstrate below that many of the results are surprisingly insensitive to the shape of *f*, depending only on the standard deviation τ.

In addition to the Gaussian assumption regarding *f*, we also assume that the stochastic process describing the evolution of relative fitnesses has reached a stationary state, which we label $p(w)$. We are unaware of analytic theory sufficient to describe $p(w)$ given a heritability function *f* and population size *N*. In the absence of such theory, we assume that $p(w)$ is the skew-Gaussian distribution (Azzalini 1985), with probability density function given by

$$p(x) = \frac{1}{\omega\pi} e^{-\frac{(x-\varepsilon)^2}{2\omega^2}} \int_{-\infty}^{\frac{\alpha(x-\varepsilon)}{\omega}} e^{-\frac{t^2}{2}} dt, \qquad (1)$$

with mean $\varepsilon + \omega\delta\sqrt{2/\pi}$, variance $\sigma^2 = \omega^2\left(1 - \frac{2\delta^2}{\pi}\right)$, and skewness $\frac{4-\pi}{2}\frac{(\delta\sqrt{2/\pi})^3}{(1-2\delta^2/\pi)^{3/2}}$, where $\delta = \frac{\alpha}{\sqrt{1+\alpha^2}}$. In our constant population size model, each individual has one offspring on average, and thus, we set the mean of $p(w)$ at unity. The two remaining parameters, describing the variance (denoted σ^2) and skewness of the distribution of fitnesses, depend on *N*, τ, and potentially the higher moments of *f*. Figure 1*c* and d compare the skew-Gaussian assumption for $p(w)$ with the steady-state distributions attained in both the discrete-sites model and the Gaussian approximation described above.

Performing the calculations below requires choosing exact values for both population variance, σ^2, and skewness parameter α for $p(w)$. In lieu of analytic results, we resort to simulation data to find the appropriate σ^2. In the results that follow, the σ^2 corresponding to a particular *N* and τ has been interpolated from simulation runs conducted for each combination of *N* and τ (the length of each run varied with *N*, but at minimum 1 million generations were simulated, with the first 5*N* generations discarded as burn-in). The distribution of population fitnesses exhibited leftward (negative) skew. Our results in general are not strongly dependent on the choice of α. Except when indicated, we use $\alpha = -2$. The resulting function closely, but not exactly, matches the actual distribution of fitnesses observed in simulation results (fig. 1*c* and *d*). Nonetheless, this choice of α yields results that are broadly consistent with those obtained in simulations over a range of parameter values.

Distribution of Ancestral Fitnesses

Henceforth, we assume that $f(w_o; w_p, \tau^2)$ is Gaussian with $\tau^2 = L\mu s^2$. We first seek to calculate the probability that an ancestor of a randomly selected individual *t* generations

in the past had fitness w. Assuming that the fitness, W_o, of some individual in question is described by probability density $\phi(w_o)$, we seek a function describing the probability that the parent of the individual had fitness w_p. Symbolically,

$$Pr\{W_p = w_p\} = \int Pr\{w_p|w_o\}\phi(w_o)\mathrm{d}w_o. \quad (2)$$

We can reverse the conditioning on the right-hand side via Bayes' theorem to obtain the following:

$$Pr\{W_p = w_p|w_o\} = \frac{Pr\{w_o|w_p\}Pr\{w_p\}}{Pr\{w_o\}}. \quad (3)$$

$Pr\{w_o|w_p\}$ is the probability that an offspring has fitness w_o conditional on parent fitness and may be expressed as the product of two quantities. First, $f(w_o;w_p,\tau^2)$ is the probability that any parent with fitness w_p could have given rise to an offspring with w_o. Because our model tracks only relative fitness, we must account for the fact that each offspring fitness is divided by the mean fitness of all offspring each generation. The expected value of this normalization factor (the mean fitness of offspring prior to normalization) may be expressed as the mean fitness of the parental generation plus the expected increase in fitness of the offspring. In our model, the mean fitness of the parents is exactly one, whereas Fisher's fundamental theorem states that the expected increase in fitness of the offspring is equal to the variance in fitness in the parental generation (Fisher 1931). At stationarity, the variance of fitnesses in the parental generation is σ^2, and therefore, the expected normalization factor is $1+\sigma^2$. Although sampling error may cause the actual value in a given generation to differ from the expected value, in the calculations below we ignore these fluctuations and assume that the factor is exactly $1+\sigma^2$. Second, the factor w_p/N accounts for the fact that fitter individuals are more likely to be a randomly chosen individual's parent. The product $w_p f(w_o;w_p,\tau^2)/N$ then describes the probability that an offspring with fitness w_o had a parent with fitness w_p.

$Pr\{w_p\}$ is the probability that a member of the previous generation with unknown reproductive success had fitness w_p and thus is given by the stationary distribution of fitnesses $p(w)$. The denominator in equation (3) is a normalizing factor that can be computed by integrating the expression in the numerator over all possible values of w_p. Thus, the full probability that a randomly chosen individual with fitness w_o had a parent with fitness w_p is

$$Pr\{W_p = w_p|w_o\} = \frac{w_p f\left(w_o;\frac{w_p}{1+\sigma^2},\tau^2\right)p(w_p)}{\int z f\left(w_o;\frac{z}{1+\sigma^2},\tau^2\right)p(z)\mathrm{d}z}. \quad (4)$$

In the current generation, we have no information regarding individual's reproductive success, thus a randomly chosen individual has fitness described by $p(w)$ and we set $\phi(w_o) = p(w_o)$ in equation (4). Performing the integrations in equation (4) results in a distribution describing the parental fitness of a randomly selected individual. Let the distribution obtained in the above procedure be $\phi_1(w)$,

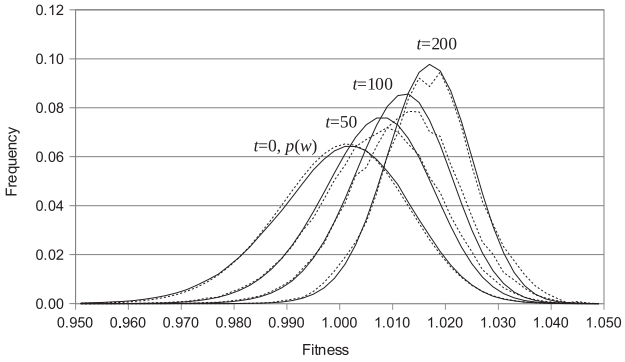

FIG. 2. Distributions of ancestral fitnesses ($\phi_t(w)$), from simulation (dashed lines) and numerical analysis (solid lines), at several time points in the past (indicated by text above each set of curves). $N = 1,000$, $\tau = 5 \times 10^{-4}$, and $\alpha = -1.25$.

with the 1 indicating ancestral fitnesses one generation in the past. Additional ancestral distributions may be obtained through the recurrence relation

$$\phi_{t+1}(w) = \int_0^\infty \frac{w f\left(w_o;\frac{w}{1+\sigma^2},\tau^2\right)p(w)}{\int z f\left(w_o;\frac{z}{1+\sigma^2},\tau^2\right)p(z)\mathrm{d}z}\phi_t(w_o)\mathrm{d}w_o. \quad (5)$$

Repeated application of equation (5) results in a series of distributions describing ancestral fitnesses at each generation in the past. Note that the distributions describe the fitness of the ancestors of a single individual, not the group of individuals ancestral to the current generation. Henceforth, we use $\phi_t(w)$ to describe the distribution resulting from t iterations, yielding the distribution of ancestral fitnesses t generations ago.

Several distributions for different time points are shown in figure 2 in which two primary phenomena are evident. First, the expectation of ancestral fitness increases with t. Intuitively, this makes sense because individuals that leave more offspring are more likely to be represented in the current generation. Similar findings have been noted by Barton and Etheridge (2004) and Wakeley (2001), who both found a tendency for lineages to migrate toward more fit states. Second, the variance in ancestral fitnesses decreases, which also follows from lineages tending toward a narrow range of states. The change in both the mean and the variance of ancestral fitness decreases with time and approaches an equilibrium. The position and scale of the distribution at equilibrium depends on both N and τ, with large populations support more variability in fitnesses for a given τ, which leads to higher ancestral fitnesses. Larger τ also yields higher ancestral fitnesses as well as a more rapid approach to this level (results not shown).

Coalescent Rate and the Distribution of Pairwise Coalescence Times

Using the distributions of ancestral fitnesses described by $\phi_t(w)$, it is possible to compute the probability that any two individuals randomly sampled from the current generation first had the same ancestor t generations ago. Some approximation is involved. In particular, we assume that the

fitnesses of the ancestors of the two individuals, say w_A and w_B, from t generations ago, are described by independent draws from $\phi_t(w)$. This may not be the case in reality, for instance, if significant correlations exist in the distribution of fitnesses within generations, then knowing w_A may influence the distribution of w_B. This concern may be relevant to small populations with large amounts of fitness variability, such that only a few individuals found the entire next generation. Additionally, because we assume that the two lineages have not yet coalesced, their true fitnesses are likely to be more different than expected under the independence assumption (see Wilkins and Wakeley 2002). Nonetheless, the approximations are accurate for the parameter combinations examined here.

Assume two individuals A and B in the same generation have fitness w_A and w_B. The probability that a randomly selected member of the previous generation is the parent of A is given by

$$\frac{1}{N}\int \frac{wf\left(w_A;\frac{w}{1+\sigma^2},\tau^2\right)}{\int zf\left(w_A;\frac{z}{1+\sigma^2},\tau^2\right)}p(z)\mathrm{d}z p(w)\mathrm{d}w = \frac{1}{N}. \quad (6)$$

The probability that any two individuals had the same parent in the previous generation is the product of the probability that a particular individual is the parent of both A and B, summed over all possible parents. Conditional on w_A and w_B, this is

$$\gamma(w_A,w_B) = \frac{1}{N}\int \frac{wf\left(w_A;\frac{w}{1+\sigma^2},\tau^2\right)}{\int zf\left(w_A;\frac{z}{1+\sigma^2},\tau^2\right)p(z)\mathrm{d}z} \times \frac{wf\left(w_B;\frac{w}{1+\sigma^2},\tau^2\right)}{\int zf\left(w_B;\frac{z}{1+\sigma^2},\tau^2\right)}p(z)\mathrm{d}z p(w)\mathrm{d}w. \quad (7)$$

If the exact values w_A and w_B are not known, but instead are drawn from a known distribution, then the marginal probability of coalescence in the previous generation may be obtained by integrating equation (7) over the distributions describing w_A and w_B. If A and B are randomly selected from the same generation, then each has fitness w with probability described by $p(w)$, and integrating over this distribution for both w_A and w_B yields the probability that two random individuals had the same parent, which is the reciprocal of the inbreeding effective population size. If there is no variance in fitness, then $p(w)$ is a delta function at $w = 1$, and the probability of shared parentage is $1/N$, as predicted by neutrality.

If individuals A and B are ancestors from t generations ago of two individuals sampled at the current generation, then each has fitness w with probability $\phi_t(w)$. Double integration of equation (7) over $\phi_t(w)$ for both w_A and w_B then yields an expression for the probability that any two individuals in the present generation shared a common ancestor t generations ago, conditional on the two lineages not coalescing by generation $t-1$. Denoting this probability λ_t, we have

$$\lambda_t = \int\int \gamma(w_A,w_B)\phi_t(w_A)\phi_t(w_B)\mathrm{d}w_A\,\mathrm{d}w_B. \quad (8)$$

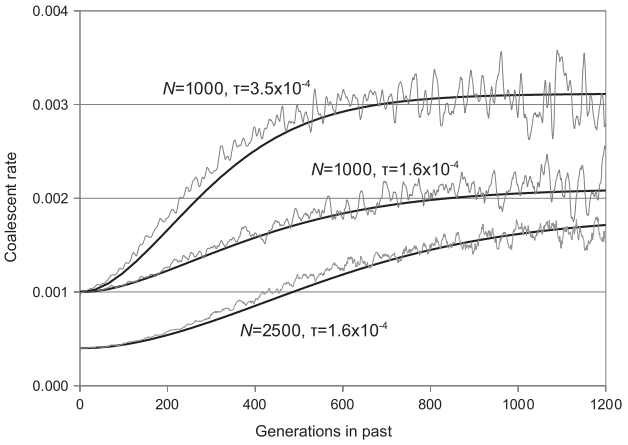

FIG. 3. Pairwise coalescent rate λ_t (eq. (8)) for several combinations of N and τ, both simulated (gray lines) and numerically calculated using the Gaussian model (black lines).

For small t, λ_t is close to the reciprocal of the inbreeding effective population size. However, λ_t increases with t in roughly linear fashion at first and then approaches an asymptote (fig. 3). Because λ_t reflects the probability of coalescence, it can be thought of as the reciprocal of a time-dependent inbreeding effective size, which decreases as one looks into the past, before eventually reaching an equilibrium size. In Seger et al. (2010), this effective size is plotted as a function of t and selection strength. If τ is very large, indicating either strong selection or high mutation rate, the approach to this equilibrium may be very fast. In this regime, simply assuming that the population has a constant effective size somewhat smaller than the census size may accurately describe the effects of selection; this is similar to the "background selection" limit proposed by Charlesworth et al. (1993). However, for smaller τ, the approach to the equilibrium is more gradual and must be taken into account to describe genealogies accurately. For a given τ, relatively large populations experience a greater increase in coalescent rate than do smaller populations (cf. $N = 2,500, \tau = 1.6 \times 10^{-4}$ and $N = 1,000, \tau = 1.6 \times 10^{-4}$; fig. 3), such that two populations of very different sizes may have similar final, "asymptotic" coalescent rates. For this reason, population size may have a less pronounced effect on the genealogies of genes under selection than on neutral genealogies.

The probability that a coalescent event between two lineages first happened on generation t is given by

$$\psi(t) = \prod_{i=1}^{t-1}(1-\lambda_i)\lambda_t. \quad (9)$$

This distribution of pairwise coalescent times is similar to the geometric distribution that describes the neutral pairwise coalescent times in discrete time models, but here, the rate λ ($1/N$ under neutrality) increases with time. As expected from the results above, increasing τ results in a substantial reduction in both the mean and the variance of the pairwise coalescent time distribution (fig. 4). Even relatively

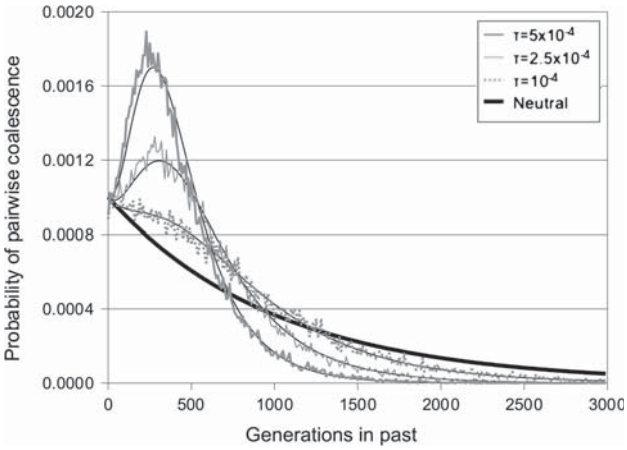

FIG. 4. Distribution of pairwise coalescence times, for $N = 1,000$ and three different values of τ, comparing numerical methods (black lines) with simulations using the Gaussian model (gray lines).

small values create a noticeable distortion when compared with the neutral expectation. For instance, at $\tau = 10^{-4}$, the mean time to coalescence is reduced from 1,000 to near 700, and the standard deviation is reduced from 1,000 to roughly 600. Greater values of τ result in considerable reductions in mean time to coalescence, for instance, at $\tau = 0.00025$, the mean time is reduced to approximately 540, with a standard deviation of 416.

Timing and Length of Coalescent Intervals

The calculation of the distribution of pairwise coalescence times says little about the structure of genealogies with more than two tips. One way to gather additional information about tree shape is to examine the distribution of lengths of different coalescent intervals. Under neutrality, these are straightforward to calculate; the expected length for an interval with j lineages is $2N/j(j-1)$. Under selection, the times are more complex because they must take into account how long ago the interval occurred because fitnesses and thus probability of coalescence change as a function of time.

Assuming a sample of size n, and population size $N \gg n$, let T_n be the time to the first coalescent event in the genealogy. In a manner similar to the pairwise case, we assume that the fitness of each lineage is an independent draw from $\phi_t(w)$, whereas noting that the true distribution of fitnesses is likely to have a greater variance (see Coalescent Rate and the Distribution of Pairwise Coalescence Times section). Under the assumption of both independent fitnesses and population size much larger than the sample size, a coalescence among the n lineages in the sample occurs on generation t with approximate probability

$$\binom{n}{2}\lambda_t. \tag{10}$$

This time-dependent probability may then be converted into a distribution using the same reasoning as for equation (9), yielding an expression for the probability distribution of T_n. The distribution for T_{n-1} may be found similarly but must be conditioned on the distribution T_n.

$$Pr\{T_{n-1} = t\} \approx \sum_{x=1}^{t-1} Pr\{T_{n-1} = t | T_n = x\} Pr\{T_n = x\}, \tag{11}$$

$$\approx \sum_{x=1}^{t-1} \prod_{i=x+1}^{t} \left(1 - \binom{n-1}{2}\lambda_i\right) \times \binom{n-1}{2}\lambda_t Pr\{T_n = x\}. \tag{12}$$

This procedure yields the total time until the second interval ends; the length of the interval may be found by subtraction. Distributions for other intervals may be found in the same manner but require iterative calculation of more recent intervals in order to obtain the distribution of starting times for the interval in question. This computation relies on the assumption that the pairwise coalescence rate λ_t is unaffected by coalescent events. In reality, the expected fitness of a lineage immediately after a coalescence (in backwards time) is higher than predicted by $\phi_t(w)$ because the ancestor had at least two offspring in one generation. Changes in fitness cause the rate of coalescence to deviate from λ_t. This deviation is likely to be transient, however, and as t increases the expected fitness distribution will approach $\phi_t(w)$ and the rate will return to λ_t. Equation (11) assumes that the return to $\phi_t(w)$ is instantaneous. Inclusion of the transient deviations in coalescence rate due to prior coalescences leads to skewed trees, a property investigated in the Seger et al. (2010).

These calculations were performed for several parameter combinations and the expectations of the resulting length distributions compared with simulation results in figure 5. Several features are evident. First, increasing population size while holding τ constant results in an ever greater distortion of the genealogy from neutral expectation, particularly near the root, where coalescence times are much reduced. Along intermediate intervals, the length of coalescent intervals are similar to those predicted under neutrality. The numerical approximation consistently predicts a somewhat smaller deviation from neutrality near the basal nodes than does the simulation. This inaccuracy seems to result from a somewhat lower equilibrium fitness resulting from the calculations than actually found in simulations; an error that decreases the coalescence rate deep in the genealogy and may result either from inaccuracy in the prior $p(w)$ or the assumption that the fitnesses are independent draws from $\phi_t(w)$. Second, the relative lengths of the one or two intervals closest to the tips are actually larger than the neutral expectation in the simulations. This reason for this remains unclear, although it may be related to the ambiguity in measuring interval lengths when more than one coalescent event occurs in a single generation. This explanation is consistent with the observation that the phenomenon decreases with population size. The distortion appears quite transient and is not likely to strongly influence patterns of nucleotide diversity.

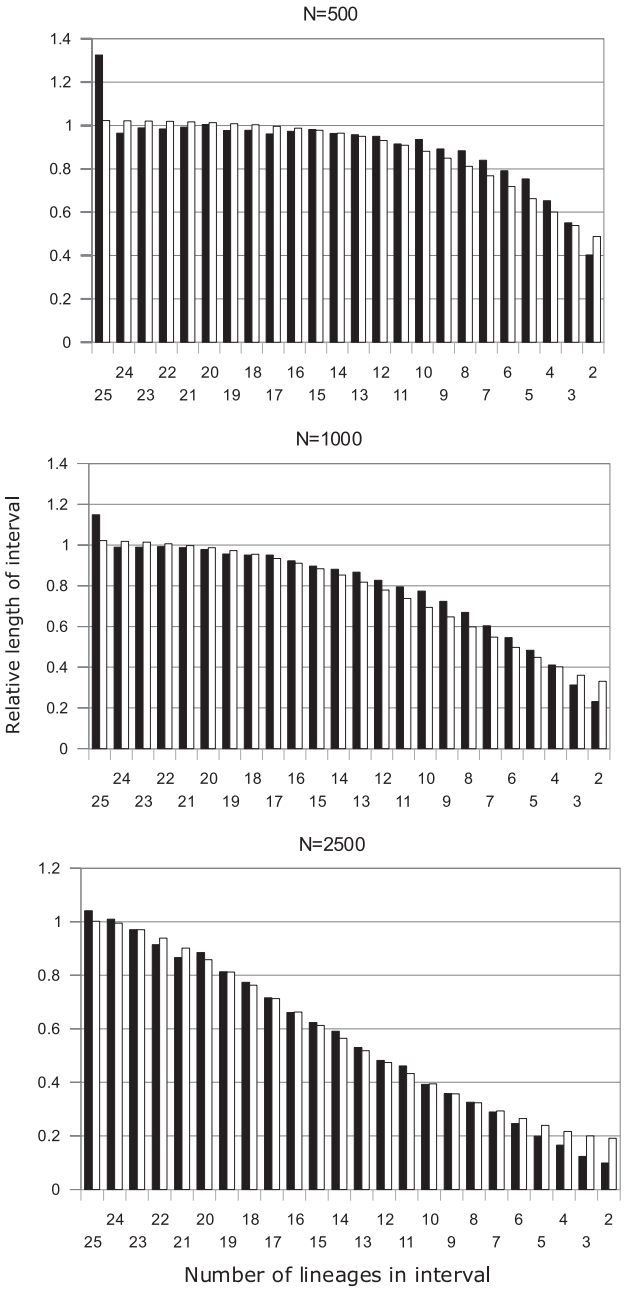

FIG. 5. Lengths of coalescent intervals relative to neutrality for a sample of size 25, $\tau = 0.0005$, and several choices of N. Black bars are simulation runs using the Gaussian model and white bars are numerical results.

The generation at which final coalescent event occurs is the TMRCA of the sample. This timing of this event describes the depth of the tree and thus affects the total tree length and the total number of mutations that have arisen in the genealogy. The distribution of this event is particularly strongly influenced by heritable variation (fig. 6), with increasing τ reducing both the mean TMRCA and its variance. A modification of the procedure described in this section may also be used to find the total length of the tree, which is given by the sum over all intervals of the number of lineages in the interval multiplied by the length of the interval.

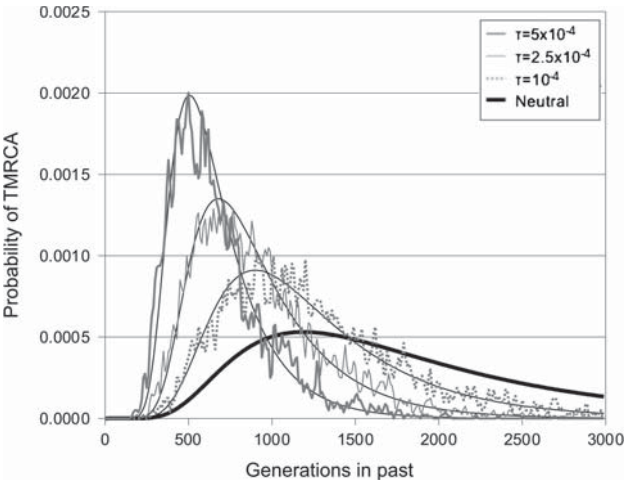

FIG. 6. The distribution of TMRCA for a sample of size 25, for $N = 1,000$, and τ as shown in figure legend, for both simulated under the Gaussian model (gray lines) and numerically calculated (black lines) results.

Results

Comparison with Discrete-Sites Models

To validate our methods, we return to the discrete-sites model introduced at the beginning of the Methods section and compare the mathematical results from the continuous model to forward simulation of the discrete-sites case. We stress that the discrete-sites model makes no assumptions regarding the form of f or the continuous nature of relative fitnesses; it is nearly identical to the model presented in Rouzine et al. (2003) (without compensatory mutations) and Seger et al. (2010). For all results shown, simulations were allowed to burn in for $10N$ generations before data recording began. After burn-in data were sampled every 1,000 generations for at least 5 million generations. We set $\tau^2 = L\mu s^2$ in all cases. For a variety of selection coefficients and population sizes, the predicted distributions of both pairwise coalescent times and the TMRCA closely match those obtained from simulation of the more complex discrete-sites model (fig. 7). For larger selection coefficients, however, such that $Ns > 10$, the continuous approximation becomes inaccurate, and the mathematical analysis predicts a greater distortion than observed in simulations (results not shown). One potential cause of this disagreement is that when selection is relatively strong, few sites are segregating, and the stationary state distribution $p(w)$ is no longer accurately represented by a continuous function. The effects of strong background selection have been investigated by other authors (Charlesworth et al. 1993; Wakeley 2008).

To more closely examine the relationship between the number of segregating sites and the accuracy of the approximation, we conducted additional simulations with $N = 1,000$, $Ns = 1.0$, and varying levels of L and μ, while maintaining the genomic mutation rate $L\mu = 0.01$ or $L\mu = 0.1$. The results in table 1 demonstrate that increasing the number of selected sites beyond 1 reduces both the pairwise

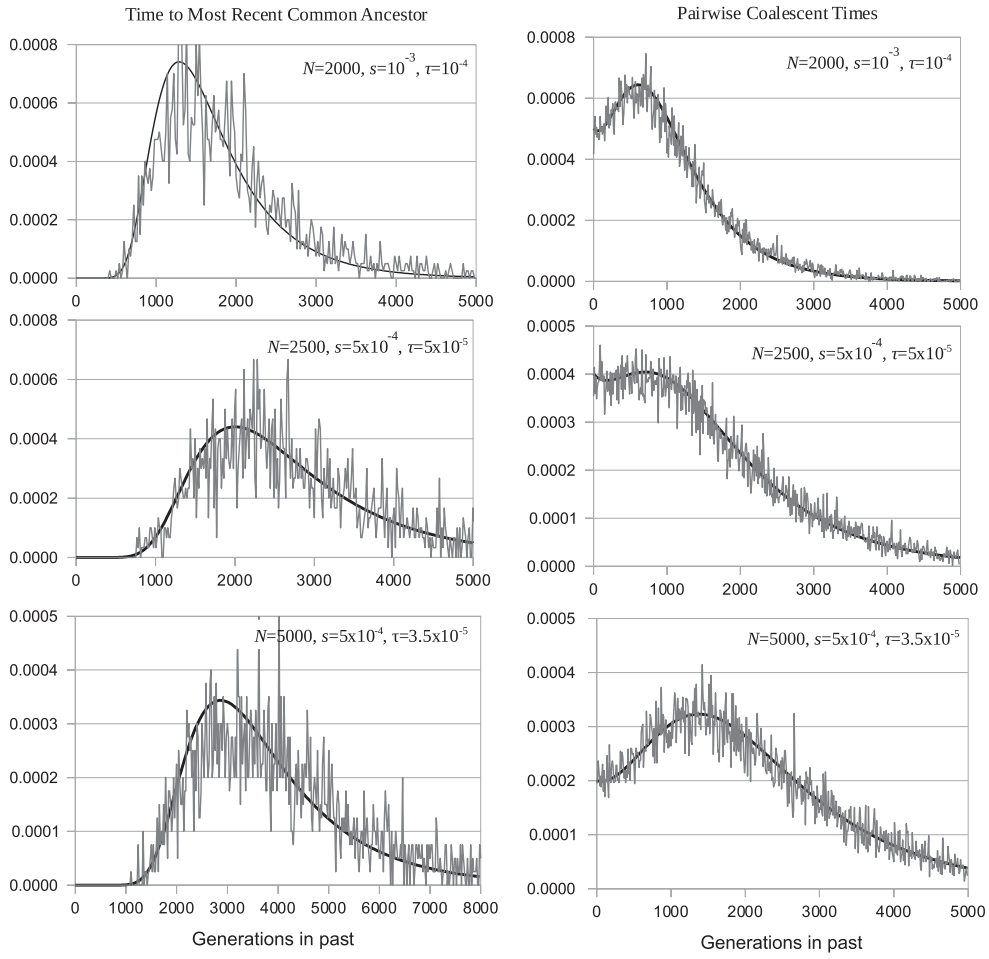

FIG. 7. Distributions of TMRCA (left column) and pairwise coalescence (right column) for simulations using the discrete-sites model of fitness heritability (thin gray lines), with numerical results (black lines). Sample size for TMRCA calculation was 25. For the discrete-sites model, $N = 1,000$, $\mu = 10^{-5}$ for $N = 2,000$, and $N = 2,500$, $\mu = 5 \times 10^{-6}$ for $N = 5,000$.

coalescent time and the TMRCA substantially and therefore that one-locus, two-allele models (equivalent to $L = 1$) should not be used to infer the effects of selection at many sites, even when the expected mutation probability $L\mu$ is held constant. As L increases, the population harbors more fitness variability, as measured by σ, with σ eventually ceasing to increase for $L > 1000$.

Robustness of Results to Other Fitness Heritability Functions

Although the results in figure 7 indicate that models with different assumptions regarding the heritability of fitness may have similar effects on genealogies provided they share the same τ, we further pursue this topic by examining different forms of the fitness heritability function f via forward simulation (fig. 8). Despite the considerable differences in the form of f, the distributions of pairwise coalescent times are indistinguishable, and the means and variances differ by only a few percent. Identical results were obtained with other population sizes and values of τ (results not shown). The similarity suggests that the higher moments of f do not greatly, or even moderately, affect genealogical structure and that the process is influenced primarily by the variance of the fitness heritability function, τ^2. One possible explanation for this result is the central limit theorem-like property of adding many small deviations, such that offspring fitness distributions are approximately Gaussian when compared with ancestors from many generations ago, regardless of the form of the deviation produced in each generation. These results suggest that analytic calculations using a Gaussian f may accurately describe genealogical relationships for a substantially larger class of models, including more traditional finite- or infinite-sites models.

Discussion

This work demonstrates that a model of coalescence involving continuously variable fitnesses can capture some of the ways in which weak selection acting at many sites distorts genealogies from their neutral expectation. We find that weak selection, on the order of $Ns \approx 1$, can significantly shorten the time taken for two lineages to reach a common ancestor and that the variance of this time is reduced by an even greater factor (fig. 4). Weak selection also distorts the shapes of larger trees by shortening the lengths of coalescent intervals near the root of the tree while leaving the

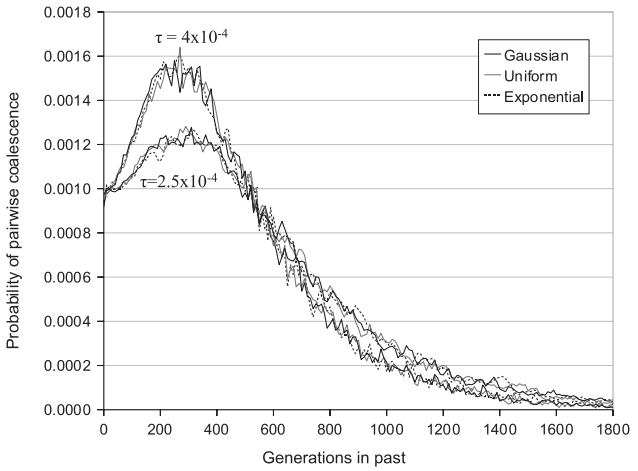

FIG. 8. Distributions of pairwise coalescence times for different choices of f and τ. $N = 1,000$ in all cases.

other intervals similar in length to the neutral expectation (fig. 5). The TMRCA of a sample may also be significantly reduced, in some cases by nearly 50% compared with the neutral expectation (fig. 6). Although our numerical methods assume an infinite number of possible fitness states, our calculations appear accurate when compared with simulations with more than a few hundred sites.

Although a number of authors have constructed models of the coalescent with selection, most have focused on the action of selection at a single locus (Golding 1997; Neuhauser and Krone 1997; Barton and Etheridge 2004; Wakeley 2008) and have largely concluded that weak selection does not significantly impact the shape of genealogical trees. Simulation studies of selection at multiple sites have reinforced this view (Przeworski et al. 1999; Williamson and Orive 2002). The analysis here offers a contrasting view and demonstrates several ways in which tree shape is distorted as a result of selection. In particular, table 1 demonstrates that as the number of selected sites increases, populations harbor a greater variance in fitnesses (σ), and this increase is commensurate with increased genealogical distortion. Among authors who have examined multiple-sites simulations, some examined only selection coefficients too small to have an impact (Przeworski et al. 1999; $Ns = 0.1$). Others found deviations similar to those here but concluded that "selection only had a moderate effect on tree statistics…consistent with single locus results" (Williamson and Orive, 2002, p. 1379). Given that selection is likely to act at many sites simultaneously in natural populations and that many mutations are likely to impact reproduction only modestly, this work suggests that many real genealogies may experience considerable distortions due to selection.

One intriguing finding of this analysis is the fundamental importance of the variance τ^2 to the exclusion of the other properties of the distribution f (fig. 8). The parameter τ reflects the extent to which offspring fitness may differ from parental fitness and thus incorporates both mutation and selection. If mutation never occurs, or if mutations have no impact on fitness, then offspring will always have the same

Table 1. Effect of L on genealogical distortion.

L[a]	τ[b]	σ[c]	Pairwise Coalescent Time[d]	TMRCA[e]
$L\mu = 0.01$				
1	0.0001	0.0005	973 (961)	1,875 (1,029)
10	0.0001	0.0014	958 (1,045)	1,857 (1,241)
100	0.00011	0.0021	746 (636)	1,400 (642)
1,000	0.00012	0.0022	714 (594)	1,320 (567)
2,500	0.00012	0.0022	711 (582)	1,286 (567)
Calc[f]	0.0001	0.0021	681 (562)	1,212 (555)
$L\mu = 0.1$				
1	0.0003	0.0005	996 (1,072)	1,924 (1,230)
10	0.00032	0.0016	1,022 (1,084)	1,976 (1,196)
100	0.00032	0.0041	657 (560)	1,213 (573)
1,000	0.00032	0.0053	488 (381)	889 (396)
2,500	0.00032	0.0055	482 (360)	880 (363)
Calc[f]	0.0003	0.0050	499 (378)	854 (368)

[a] Number of sites under selection.
[b] Standard deviation of the difference parent–offspring fitness.
[c] Standard deviation in population fitness distribution.
[d] Average pairwise coalescence time in generations, with standard deviation in parentheses.
[e] Time to most recent common ancestor, with standard deviation in parentheses.
[f] Results calculated using the numerical method with $\tau = L\mu_s^2$, numbers in parenthesis are standard deviations of the calculated distributions.

fitness as their parents, $\tau = 0$, and genealogies conform to the neutral coalescent. Conversely, if mutations are frequent and their impact on fitness is large, offspring fitness may differ greatly from parental fitness, τ will be very large and coalescences will happen much more rapidly than predicted under neutrality.

Why variance, but not, for instance, skewness, should be the primary determining factor of coalescence time remains unclear, but in many ways, it is fortunate. For instance, it suggests that very different mutational models may still have similar effects on genealogies, provided that they produce similar variances in fitness heritability. This finding may be particularly important when comparing analytic results with empirical data because in any real data set, the distribution of mutation probabilities and selection coefficients across a sequence are likely to be unknown. Nonetheless, the influence of a complicated mutational and selective regime on genealogical structure may be similar to the Gaussian model presented here, provided that τ^2 is the same. This result is supported by investigations of a finite-sites model of mutation and selection (fig. 7), which produces results very similar to those predicted by the continuous coalescent (as long as L is large enough), despite very different assumptions regarding the mutational model.

Although τ is presented here as a parameter of the model, τ may be calculated for any model in which the reproductive success of parents and offspring may be compared. Specifically, τ is the standard deviation of the differences between parent and offspring expected reproductive successes. τ may be estimated by tabulating the difference in actual reproductive success between a parent and its offspring over many parent–offspring pairs and then calculating the variance among these differences. For haploid organisms with nonrecombining genomes, it is possible that

τ may be calculated in laboratory studies. However, for organisms with recombining chromosomes, empirical assessment of τ is likely to be more difficult, unless a single polymorphism can be tracked. Nonetheless, it may be possible to estimate τ from a genealogy reconstructed for a particular genomic region.

One advantage of the approach presented here is that it allows for the likelihood of a genealogy to be calculated given a particular population size and τ. Using a genealogy sampler such as LAMARC (Kuhner, 2006) or BEAST (Drummond and Rambaut, 2007), it may then be possible to estimate the true values of these parameters using sequence data from natural populations. Such an approach would allow for estimation of the amount of heritable fitness variation produced at unlinked loci in a single generation, facilitating a test for selection based on deviations in genealogical structure. We consider the feasibility of such an approach in a forthcoming publication.

The work presented here is similar to models of coalescence in continuous habitats (Barton and Wilson 1995; Wilkins and Wakeley 2002; Wilkins 2004). These models typically assume strict population regulation, such that population density is distributed uniformly across the habitat as well as Gaussian dispersal of offspring. Under these assumptions and using a diffusion approximation, Wilkins and Wakeley (2002) found an analytic expression for the full distribution of pairwise coalescence times, conditional on the starting location of the samples, and the variance of the dispersal function. The analysis in this paper may be seen as extending Wilkins and Wakeley (2002) model to include differential reproductive success along the habitat in a linear manner and assuming that population density is given by the skew-Gaussian distribution. One key difference, however, is that the model here rescales "space" each generation so that the mean "location" (fitness) is unity.

In order for the work here to be fully analytic, a function must be found that describes the steady-state distribution of fitnesses in a population with a particular mutation and selection model. Deriving this expression may require solving a functional equation relating the distribution of offspring fitnesses to the distribution of parental fitnesses. A potential alternative is to find an expression for the moments of the steady-state distribution by solving a system of equations relating each moment of the parental distribution to the offspring distribution. Such an approach requires a method of moment closure because each moment of the parental distribution affects a different moment in the offspring distribution (for instance, the mean of the offspring distribution is governed by the variance of the parental distribution). Additionally, as genealogical shape appears to be fairly sensitive to changes in the population variance (σ^2), any approximations made must be quite accurate over the appropriate range of parameter values.

This analysis calls into question some techniques used to infer past population dynamics. Specifically, several related techniques have been proposed to infer population growth rates and historical sizes based upon analysis of the distribution of coalescent intervals (Kuhner et al. 1998; Pybus et al. 2000; Strimmer and Pybus 2001; Minin et al. 2008). As demonstrated above, however, weak selection may produce a systematic distortion of the intervals, such that basal intervals are much shorter than expected. Performing a skyride analysis or estimating the growth rate of a population from loci experiencing moderate selection at multiple-sites selection produces a strong signal of population expansion, when in fact population size (and the population's inbreeding effective size) has remained constant (results not shown). Minin et al. (2008) analyzed the Egyptian hepatitis C virus (HCV), for example, and found a marked population expansion. However, because HCV has a high mutation rate and a single nonsegmented genome with high gene density, it seems unlikely that any region will be free from the effects of selection. Combining a skyline or growth rate analysis with the techniques presented here in a maximumlikelihood context may allow for a more robust estimate of historical population sizes and selection parameters.

Acknowledgments

We would like to thank Mary Kuhner, Joe Felsenstein, Jon Wilkins, and two anonymous reviewers for their helpful comments on a previous version of the manuscript. Financial support was provided by National Science Foundation grant DBI-0906018 to B.D.O.

Appendix

Here, we derive the mean and the variance of fitness heritability function f from the discrete-sites model. We begin by considering the distribution of the number, say J, of mutated sites in an individual conditional on the individual's parent having exactly K mutated sites. Let X be the number of forward mutations and Y be the number of back mutations. Because each site mutates independently with probability μ, X and Y are conditionally independent binomial random variables with index $L-K$ and K, respectively, each with parameter μ. Let $Z = X - Y$ be the change in the number of mutations from parent to offspring, so that $J = K + Z$. In the case where L and K are large and μ is small, and X and Y are approximated well by Poisson random variables with rate $(L-K)\mu$ and $K\mu$. The large K assumption is satisfied with the selection coefficient $s \approx 1/N$, the regime we consider here. Z is then Skellam-distributed (Skellam 1946), with mean $(L-K)\mu - K\mu = \mu(L-2K)$ and $\Pr\{J = j\} = \Pr\{Z = j - K\}$.

The probability density f is defined as the difference in relative fitness between parents and offspring, conditional on parent fitness. To find the mean and variance of f, we first denote the relative fitness of a given individual w_o. If the individual has J mutations, then $w_o = e^{-sJ}/\overline{w}$, where $\overline{w}$ is the mean absolute fitness of all individuals in the same generation as the selected individual. Similarly, the relative fitness of the chosen individual's parent is $w_p = e^{-sK}/\overline{w}$. If the population size is large and at stationary state, then $\overline{w}$ is not likely to change considerably from one generation to the next, and for these calculations, we assume it is constant.

The expectation of f is then $E[W_o - w_p|w_p] = E[W_o|w_p] - w_p$, where

$$E[W_o|w_p] = E\left[\frac{e^{-sJ}}{\overline{w}}\,|w_p\right] = \frac{e^{-sK}}{\overline{w}}E\left[e^{-sZ}\,|w_p\right]. \quad (13)$$

$E[e^{-sZ}|w_p]$ is the moment-generating function of the Skellam distribution, which is $e^{-L\mu+\mu(L-K)e^{-s}+\mu K e^{s}}$. To first order in s and μ, equation (13) simplifies to

$$E[W_o|w_p] = \frac{e^{-sK}}{\overline{w}}L\mu s, \quad (14)$$

which is equivalent to $w_p L\mu s$.

The variance of f, τ^2, can be found using a similar procedure

$$\begin{aligned}\mathrm{Var}[W_o|K] &= E[W_o^2|K] - E[W_o|K]^2 \\ &= \frac{e^{-2sK}}{\overline{w}^2}(E[(e^{-sZ})^2|K] - E[e^{-sZ}|K]^2). \quad (15)\end{aligned}$$

Again using the moment-generating function of the Skellam distribution and dropping terms of order s^2, μ^2, and $s\mu$ or higher, we find that $\tau^2 \approx \left(\frac{e^{-sK}}{\overline{w}}\right)^2 L\mu s^2$. Finally, because $E[e^{-sK}]$ is equal to the expected mean fitness of all the individuals in a given generation, $\left(\frac{e^{-sK}}{\overline{w}}\right)^2 \approx 1$, leaving $\tau^2 \approx L\mu s^2$, a value independent of both the parental state (K) and the mean fitness of any generation.

References

Azzalini A. 1985. A class of distributions which includes the normal ones. *Scand J Stat.* 12:171–178.

Barton N, Etheridge AM. 2004. The effect of selection on genealogies. *Genetics* 166:1115–1131.

Barton N, Etheridge AM, Sturm AK. 2004. Coalescence in a random background. *Ann Appl Probab.* 14:754–778.

Barton N, Navarro A. 2002. Extending the coalescent to multilocus systems: the case of balancing selection. *Genet Res.* 79:129–139.

Barton NH, Wilson I. 1996. Genealogies and geography. *Philos Transact B Biol Sci.* 349:49–59.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–1303.

Comeron, JM, Kreitman M. 2002. Population, evolutionary, and genomic consequences of interference selection. *Genetics* 161:389–410.

Comeron JM, Williford A, Kliman RM. 2008. The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. Heredity 100:19–31.

Coop G, Griffiths R. 2004. Ancestral inference on gene trees under selection. *Theor Popul Biol.* 66(3):219–232.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 7:214.

Fisher RA. 1931. The genetical theory of natural selection. Oxford: Clarendon Press.

Golding B. 1997. The effect of purifying selection on genealogies. In: Donnelly P, Tavare S, editors. Progress in population genetics and human evolution. New York: Springer-Verlag.

Hudson R, Kaplan N. 1994. Gene trees with background selection. In: Golding B, editor. Non-neutral evolution, theories and molecular data. New York: Chapman & Hall.

Hudson R, Kaplan N. 1995. Deleterious background selection with recombination. *Genetics* 141:1605–1617.

Kaplan N, Darden T, Hudson RB. 1988. The coalescent process in models with selection. *Genetics* 120:819–829.

Kingman JFC. 1982a. Exchangeability and the evolution of large populations. In: Koch G, Spizzino F, editors. Exchangeability in probability and statistics. Amsterdam: North-Holland.

Kingman JFC. 1982b. On the genealogy of large populations. *J. Appl. Probab.* 19A:27–43.

Kingman JFC. 1982c. The coalescent. *Stochastic Processes Appl.* 13:235–248.

Krone SM, Neuhauser C. 1997. Ancestral processes with selection. *Theor Popul Biol.* 51:210–237.

Kuhner MK. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* 22(6): 768–770.

Kuhner MK, Yamato J, Felsenstein J. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* 149(1):429–434.

Maia L, Colato A, Fontanari JF. 2004. Effect of selection on the topology of genealogical trees. *J Theor Biol.* 226:315–320.

McVean GAT, Charlesworth B. 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* 155:929–944.

Minin V, Bloomquist EW, Suchard MA. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol.* 25(7):1459–1471.

Navarro A, Barton N. 2002. The effects of multilocus balancing selection on neutral variability. *Genetics* 161:849–863.

Neuhauser C, Krone SM. 1997. The genealogy of samples in models with selection. *Genetics* 145:519–534.

Nordborg M. 1997. Structured coalescent processes on different timescales. *Genetics* 146:1501–1514.

Przeworski M, Charlesworth B, Wall JD. 1999. Genealogies and weak purifying selection. *Mol Biol Evol.* 16(2):246–252.

Pybus O, Rambaut A, Harvey P. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155:1429–1437.

Rouzine IM, Wakeley J, Coffin JM. 2003. The solitary wave of asexual evolution. *Proc Natl Acad Sci U S A.* 100(2):587–592.

Seger J, Smith WA, Perry JJ, Hunn J, Kaliszewska ZA, Sala LL, Pozzi L, Rowntree V, Adler FR. 2010. Gene genealogies strongly distorted by weakly interfering mutations in constant environments. *Genetics* 184:529–545.

Skellam JG. 1946. The frequency distribution of the difference between two Poisson variates belonging to different populations. *J Roy Stat Soc Ser A.* 109(3):296.

Slade PF. 2000. Simulation of selected geneaologies. *Theor Popul Biol.* 57:35–49.

Strimmer K, Pybus O. 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol Biol Evol.* 18:2298–2305.

Wakeley J. 2001. The coalescent in an island model of subdivision with variation among demes. *Theor Popul Biol.* 59:133–144.

Wakeley J. 2008. Conditional gene genealogies under strong purifying selection. *Mol Biol Evol.* 25(12):2615–2626.

Wilkins JF. 2004. A separation of timescales approach to the coalescent in a continuous population. *Genetics* 168:2227–2244.

Wilkins JF, Wakeley J. 2002. The coalescent in a continuous, finite, linear population. *Genetics* 161:873–888.

Williamson S, Orive ME. 2002. The genealogy of a sequence subject to purifying selection at multiple sites. *Mol Biol Evol.* 19(8): 1376–1384.