Distortions in Genealogies Due to Purifying Selection

Lauren E. Nicolaisen[1,2,3] and Michael M. Desai*[,1,2]

[1]Department of Organismic and Evolutionary Biology, Harvard University

[2]Department of Physics, Harvard University

[3]FAS Center for Systems Biology, Harvard University

***Corresponding author:** E-mail: mdesai@oeb.harvard.edu.

Associate editor: Rasmus Nielsen

Abstract

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. A structured coalescent method has been used to describe this effect, and forms the basis for numerical methods and simulations. In this study, 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.

Key words: coalescent, purifying selection, genealogies, linkage.

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 (Krone and Neuhauser 1997; Neuhauser and Krone 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) used 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, although 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 article, we propose an approximation that 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 effective population size result, because 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 simpler (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 by 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, whereas 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 effective 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 pre-existing 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).

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

$$h_k = \frac{\left(\frac{U_d}{s}\right)^k e^{-U_d/s}}{k!}. \quad (1)$$

This is illustrated in figure 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 backward 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 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

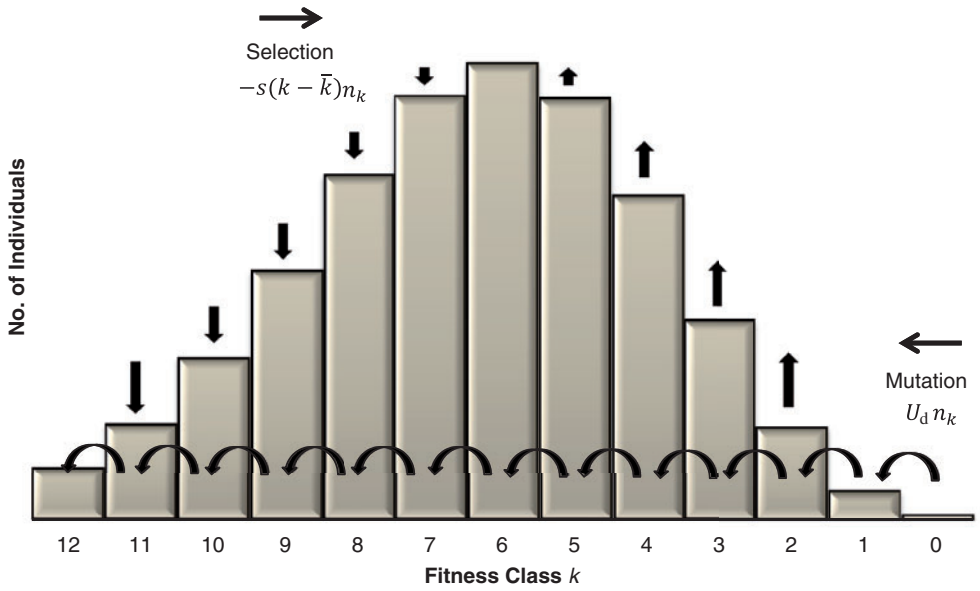

FIG. 1. Schematic depiction of mutation-selection balance. Deleterious mutations decrease the mean fitness of the population, whereas selection favors more-fit individuals. At steady state, a balance between these two effects is reached.

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:

$$\text{Rates of events}: \begin{cases} \text{Neutral mutation} & U_n \\ \text{Deleterious mutation} & sk \\ \text{Coalescence} & \frac{1}{Nh_k} \end{cases} \quad (2)$$

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 (toward higher fitness looking backward 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, for the fluctuations in fitness class k to be negligible, we require that the magnitude of selection and mutation be large compared with the size of the class. This implies that our approximation will be reasonable provided $Nh_k sk \gg 1$. In a later section, 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.

Analysis

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

$$P_1(k_i \to k_f|t) = \int \delta(t - \sum t_j) \prod_{j=0}^{j=k_i-k_f-1} s(k_i - j)e^{-s(k_i-j)t_j} dt_j \quad (3)$$

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,

$$P(k_i \to k_f|t) = e^{-sk_i t}\left(e^{st} - 1\right)^{k_i-k_f}\binom{k_i}{k_f}. \quad (4)$$

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,

$$P_{k_f}(t) = \sum_{k_i=k_f}^{\infty} h_{k_i} P(k_i \to k_f | t) = \frac{\left(\frac{U_d}{s} e^{-st}\right)^{k_f} e^{-\frac{U_d}{s} e^{-st}}}{k_f!}. \quad (5)$$

This is the ancestral fitness distribution of a randomly sampled individual; we illustrate it in figure 2. We note that, similar to 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 $t \to \infty$, 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, $\mathrm{Var}[k(t)] = \frac{U_d}{s} e^{-st}$. 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 discussed later.

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 with 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 discussed later, 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.

The Independent Lineage Approximation

The ancestral fitness distribution is defined for a single individual, moving through the distribution according to the probabilities described in equation (2). 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 $Ne^{-U_d/s}$, whereas deleterious mutation events through the distribution occur on a time scale of $\frac{1}{sk}$. Thus, we can approximate lineages as independent provided $Nse^{-U_d/s} \gg 1$.

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 toward 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,

$$P_n(t) = \frac{\binom{n}{2}}{N_e(t)}, \quad (6)$$

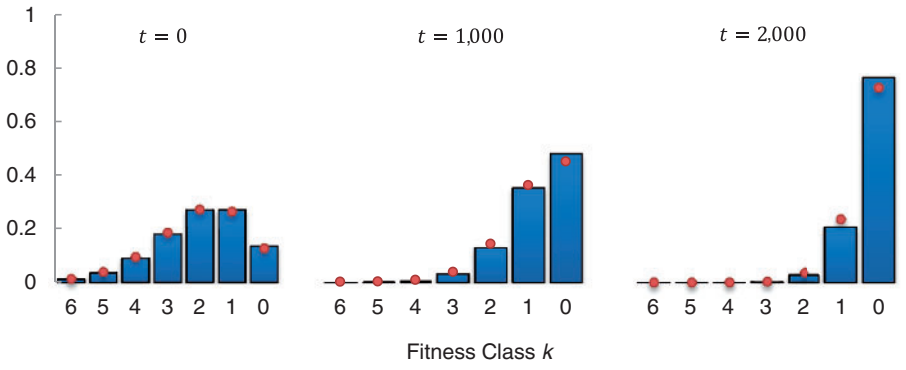

FIG. 2. The ancestral fitness distribution, shown for times $t = 0$, $t = 1{,}000$, and $t = 2{,}000$ before the present. The bars are the theoretical result, and the circles are simulations. In this plot, $\frac{U_d}{s} = 2$, $s = 0.001$, and $N = 10^5$. The present population is in mutation-selection balance. As time recedes into the past, the ancestral lineages shift toward higher fitness. As $t \to \infty$, all individuals eventually return to the zero class.

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,

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

This gives

$$N_e(t) = Ne^{-\frac{U_d}{s}(1-e^{-st})^2}. \quad (7)$$

Similarly, for arbitrary sample size, we have:

$$\frac{\binom{n}{2}}{N_e(t)} = \sum_{k=0}^{\infty} \frac{1}{Nh_k} \left[\sum_{i=2}^{n} \binom{n}{i} \binom{i}{2} P_k^i \left(\sum_{k' \neq k} p_{k'} \right)^{n-i} \right] \quad (8)$$

$$\sum_{k=0}^{\infty} \frac{\binom{n}{2}}{Nh_k} \left[\sum_{i=0}^{n-2} \binom{n-2}{i} P_k^{i+2} (1-P_k)^{n-i-2} \right]. \quad (9)$$

Using the binomial expansion:

$$(a+b)^n = \sum_{i=0}^{n} a^i b^{n-i} \binom{n}{i}, \quad (10)$$

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

$$N_e(t) = Ne^{-\frac{U_d}{s}(1-e^{-st})^2}. \quad (11)$$

Thus, we see that there is a simple $N_e(t)$ that describes any size sample. In figure 3, we illustrate our analytical prediction for $N_e(t)$ and compare it with simulation results. 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, as 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.

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, backward in time, depends on the current class of an individual. An individual at the mean fitness $\bar{k}$ has deleterious mutation rate $s\bar{k} = U_d$, as expected.

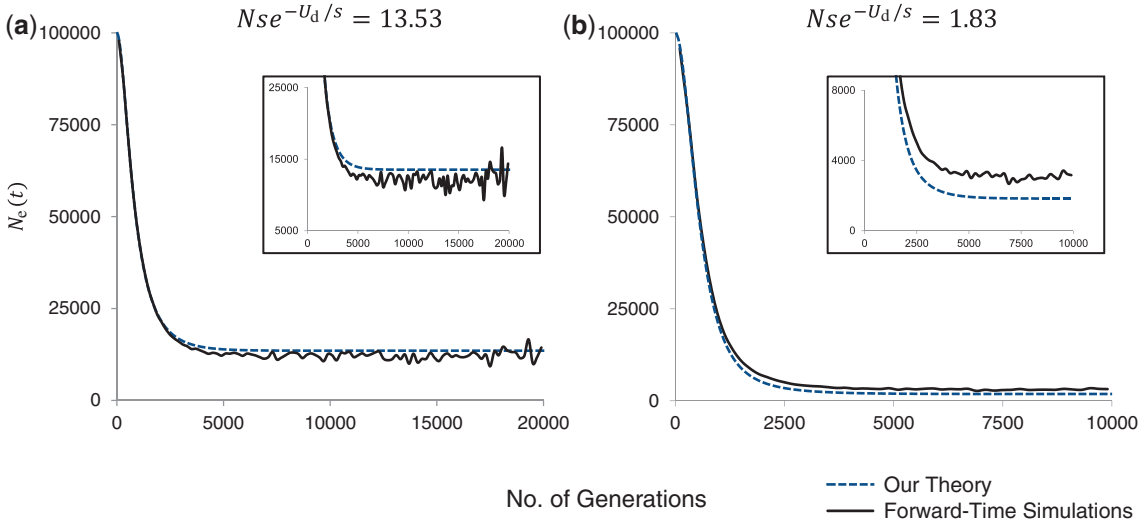

FIG. 3. Effective population size as a function of time, for **(a)** $\frac{U_d}{s} = 2$ and **(b)** $\frac{U_d}{s} = 4$. In both cases, $N = 10^5$, $U_n = U_d$, and $s = 10^{-3}$. The effective population size begins at N but undergoes a transition to a long-term rate of approximately $Ne^{-\frac{U_d}{s}}$.

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 on 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 toward 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, backward in time, at rate sk, and neutral mutations at rate U_n. Therefore, we have

$$U_e(t) = \sum_{k=0}^{\infty} P_k(t)(U_n + sk) \tag{12}$$

$$U_n + U_d e^{-st}. \tag{13}$$

In figure 4, we illustrate our prediction for $U_e(t)$ and compare it with 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 backward 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.

Simulations

We performed forward-time Wright–Fisher simulations to confirm the validity of our results. In each generation, a new set of individuals was 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 and 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 $N = 10^5$, $s = 10^{-3}$, and $\frac{U_d}{s} = 2$ Muller's ratchet proceeded between 0 and 1 times in 200,000 generations. The simulations were repeated at least 6,000 times, and the results were averaged over trials.

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 earlier. 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

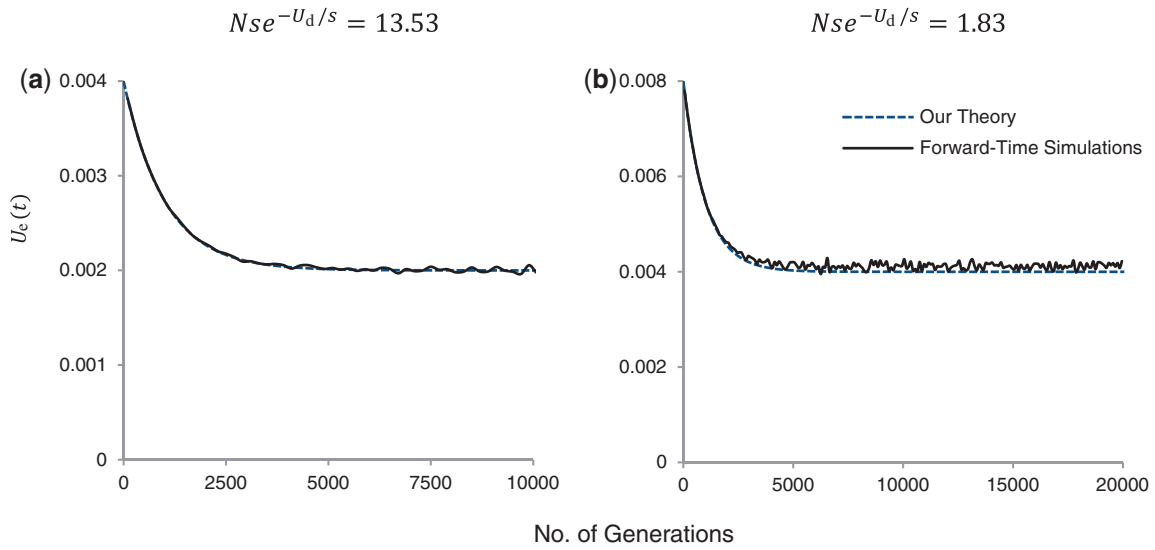

FIG. 4. Effective mutation rate as a function of time, for **(a)** $U_d = 0.002$ and **(b)** $U_d = 0.004$. In both cases, $N = 10^5$, $U_n = U_d$, and $s = 10^{-3}$. The effective mutation rate begins at the instantaneous mutation rate, $U_n + U_d$ but undergoes a transition to a long-term rate of U_n. The transition is exponentially decreasing with rate given by the selection coefficient, s.

turn used to calculate coalescent times. Other work by Seger et al. (2010) calculated a time-dependent effective population size by building on the simulated structured coalescent approach of Hudson and Kaplan (1994) and Gordo et al. (2002). Our results are also based on the framework of Hudson and Kaplan (1994) and should therefore be analogous to those earlier. Our work herein 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 (Desai et al. 2012; Walczak 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

$$\Psi_n(t) = \frac{\binom{n}{2}}{N_e(t)} e^{-\binom{n}{2}\int_0^t \frac{1}{N_e(t')}dt'}. \quad (14)$$

In figure 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 with 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 nonzero peak in the 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

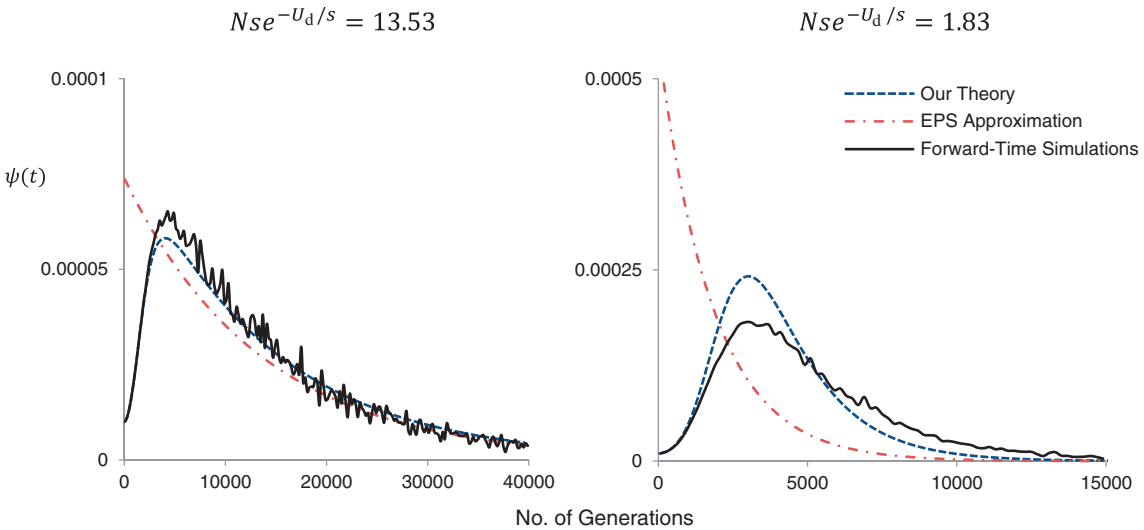

Fig. 5. Coalescence probabilities as a function of time for a sample of size two, for **(a)** $\frac{U_d}{s} = 2$ and **(b)** $\frac{U_d}{s} = 4$. In both cases, $N = 10^5$, $U_n = U_d$, and $s = 10^{-3}$. In the effective population size (EPS) approximation, the per-generation coalescence probability 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.

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

$$P_n(m|t, t_0) = \frac{\left[\int_{t_0}^{t} nU(t')dt'\right]^m}{m!} \exp\left[-\int_{t_0}^{t} nU(t')dt'\right]. \quad (15)$$

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 equation (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 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

$$P(\Pi = \pi) = \int_0^\infty \Psi_2(t) P_2(\pi|t) dt. \quad (16)$$

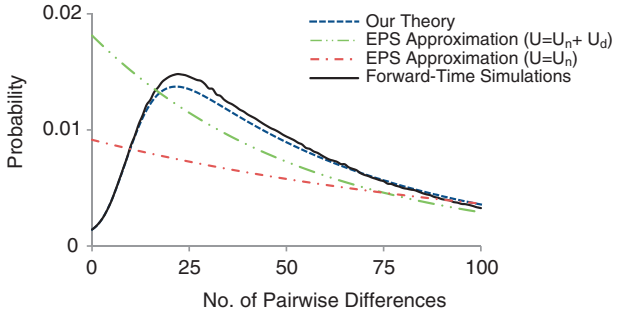

FIG. 6. Number of pairwise differences between two individuals, for $U_d = 0.002$, $N = 10^5$, $U_n = U_d$, and $s = 10^{-3}$. We compare our theoretical result with forward-time simulations. For reference, we include the effective population size (EPS) approximation for both $U = U_n + U_d$ and $U = U_n$.

We compare this theoretical result with forward-time simulations in figure 6. More complicated statistics can be calculated in an analogous manner.

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 pre-existing 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 figure 7, we compare these results with forward-time simulations of a population undergoing strong purifying selection. 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 figure 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

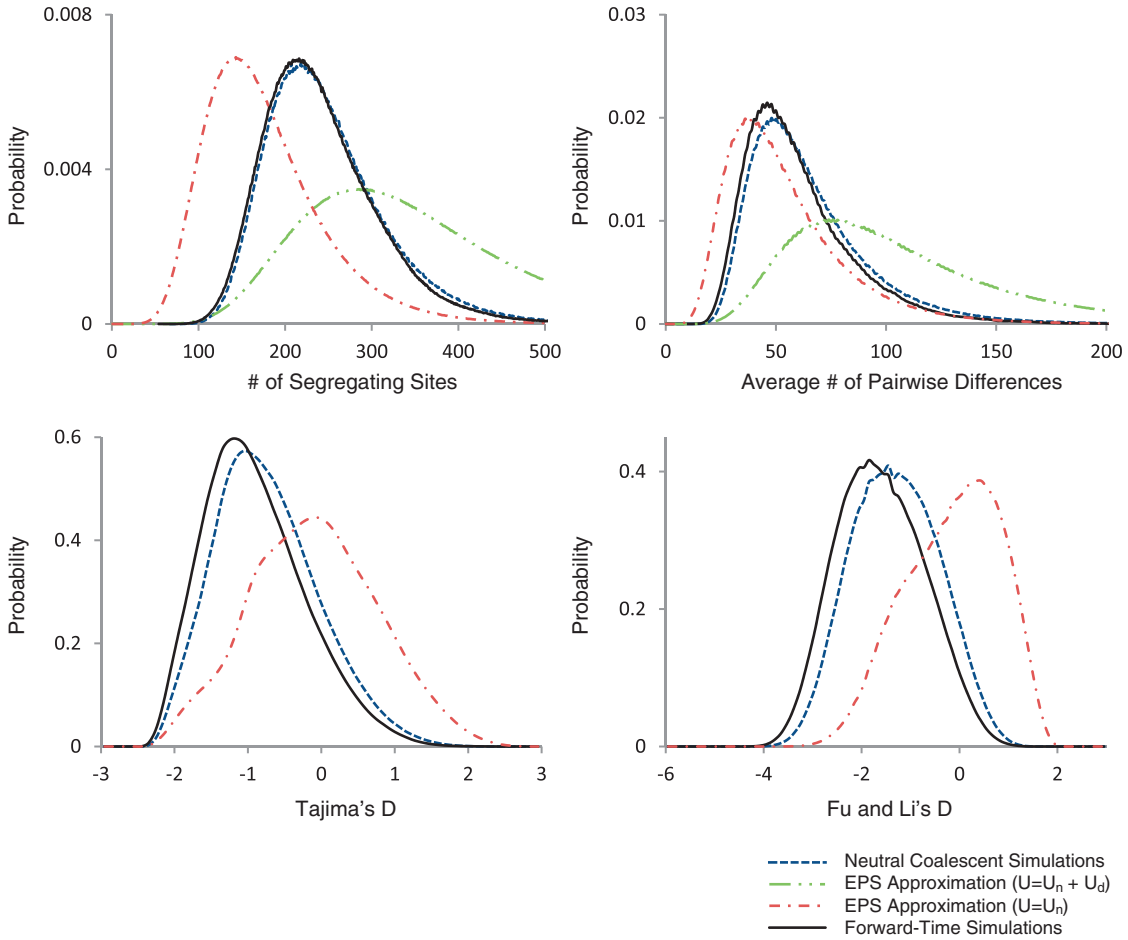

Fig. 7. Statistics for a sample of size 15, for $U_d = 0.002$, $N = 10^5$, $U_n = U_d$, and $s = 10^{-3}$. We compare neutral coalescent simulations, using our $N(t)$ and $U(t)$, with forward-time simulations under purifying selection. For reference, we include the effective population size (EPS) approximation.

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 pre-existing neutral coalescent-based methods of 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

$$\lambda(t) = 1 - \alpha(1 - e^{-\beta t}). \tag{17}$$

In comparison, our analysis shows that this function should be

$$\frac{U(t)}{U(0)} = 1 - \frac{U_d}{U_n + U_d}(1 - e^{-st}). \tag{18}$$

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 nonuniform distribution of mutations and 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 divergence occurs, with recent divergence events implying a larger mutation rate than more ancient divergence 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 mitochondrial DNA 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 to 5.0 million years. 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. 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 $\frac{U_d}{s} \approx \log 100 \approx 4.6$ and selection coefficient of $s \approx \frac{1}{timescale} \sim 5 \times 10^{-6}$.

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 $\approx 10^5$ 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 earlier, 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 nonrecombining 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 to avoid misleading results.

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 toward the more recent branches.

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

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

$$P(k_\mathrm{i} \to k_\mathrm{f}|t) = \frac{1}{sk_\mathrm{f}} \int \delta(t - \sum t_j) \prod_{j=0}^{j=k_\mathrm{i}-k_\mathrm{f}} s(k_\mathrm{i} - j) e^{-s(k_\mathrm{i}-j)t_j} dt_j$$

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

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

We can use the fact that

$$\prod_{j=0}^{k_\mathrm{i}-k_\mathrm{f}-1} (k_\mathrm{i} - j) = (k_\mathrm{i})(k_\mathrm{i} - 1) \ldots (k_\mathrm{f} + 1) = \frac{k_\mathrm{i}!}{k_\mathrm{f}!}$$

and

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

to write

$$P(k_\mathrm{i} \to k_\mathrm{f}|t) = \sum_{i=0}^{k_\mathrm{i}-k_\mathrm{f}} (-1)^{k_\mathrm{i}-k_\mathrm{f}-i} e^{-s(k_\mathrm{i}-i)t} \binom{k_\mathrm{i} - k_\mathrm{f}}{i} \binom{k_\mathrm{i}}{k_\mathrm{f}} \quad (19)$$

$$P(k_\mathrm{i} \to k_\mathrm{f}|t) = e^{-sk_\mathrm{i}t} (-1)^{k_\mathrm{i}-k_\mathrm{f}} \binom{k_\mathrm{i}}{k_\mathrm{f}} \sum_{i=0}^{k_\mathrm{i}-k_\mathrm{f}} (-e^{st})^i \binom{k_\mathrm{i} - k_\mathrm{f}}{i}. \quad (20)$$

Using the binomial equation:

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

and identifying $x = -e^{st}$ and $n = k_\mathrm{i} - k_\mathrm{f}$, this becomes

$$P(k_\mathrm{i} \to k_\mathrm{f}|t) = e^{-sk_\mathrm{i}t} (e^{st} - 1)^{k_\mathrm{i}-k_\mathrm{f}} \binom{k_\mathrm{i}}{k_\mathrm{f}}, \quad (21)$$

as claimed in the main text.

Acknowledgments

The authors 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. L.E.N was 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.

References

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

Burridge C, Craw D, Fletcher D, Waters J. 2008. Geological dates and molecular rates: Fish DNA sheds light on time dependency. *Mol Biol Evol.* 25:624.

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.

Desai MM, Nicolaisen LE, Walczak AM, Plotkin JB. 2012. The structure of allelic diversity in the presence of purifying selection. *Theor Popul Biol.* 81:144–157.

Eyre-Walker A, Keightley P. 1999. High genomic deleterious mutation rates in hominids. *Nature* 397:344–347.

Fay J, Wyckoff G, Wu C. 2001. Positive and negative selection on the human genome. *Genetics* 158:1227.

Gordo I, Navarro A, Charlesworth B. 2002. Muller's ratchet and the pattern of variation at a neutral locus. *Genetics* 161:835–848.

Hahn MW. 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 Popul Biol.* 14:251–267.

Hermisson J, Redner O, Wagner H, Baake E. 2002. Mutation-selection balance: ancestry, load, and maximum principle. *Theor Popul Biol.* 62: 9–46.

Ho S, Phillips M, Cooper A, Drummond A. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol.* 22:1561.

Ho S, Lanfear R, Bromham L, Phillips M, Soubrier J, Rodrigo A, Cooper A. 2011. Time-dependent rates of molecular evolution. *Mol Ecol* 20: 3087–3101.

Hudson R, Kaplan N. 1994. Gene trees with background selection. In: Golding B, editor. Non-neutral evolution: theories and molecular data. New York: Chapman and Hall. p. 140–153.

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

Kimura M, Maruyama T. 1966. The mutational load with epistatic gene interactions in fitness. *Genetics* 54:1337.

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

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.

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

Nielsen R, Weinreich DM. 1999. The age of nonsynonymous and synonymous mutations in animal mtDNA and implications for the mildly deleterious theory. *Genetics* 153:497–506.

O'Fallon BD. 2010. A method to correct for the effects of purifying selection on genealogical inference. *Mol Biol Evol.* 27:2406.

O'Fallon BD, Seger J, Adler FR. 2010. A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. *Mol Biol Evol.* 27:1162–1172.

Penny D. 2005. Relativity for molecular clocks. *Nature* 436:183.

Przeworski M, Charlesworth B, Wall J. 1999. Genealogies and weak purifying selection. *Mole Biol Evol.* 16:246–252.

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

Wakeley J. 2009. Coalescent theory, an introduction. Greenwood Village (CO): Roberts and Company.

Wakeley J, Hey J. 1997. Estimating ancestral population parameters. *Genetics* 145:847.

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

Weir J, Schluter D. 2008. Calibrating the avian molecular clock. *Mol Ecol* 17:2321–2328.

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

Woodhams M. 2006. Can deleterious mutations explain the time dependency of molecular rate estimates? *Mol Biol Evol.* 23:2271.

Zemlak T, Habit E, Walde S, Carrea C, Ruzzante D. 2010. Surviving historical patagonian landscapes and climate: molecular insights from galaxias maculatus. *BMC Evol Biol.* 10:67.