Title

How to infer relative fitness from a sample of genomic sequences.

Journal

Genetics, 197(3)

Authors

Dayarian, Adel
Shraiman, Boris

Publication Date

2014-07-01

DOI

10.1534/genetics.113.160986

Peer reviewed

How to Infer Relative Fitness from a Sample of Genomic Sequences

Adel Dayarian* and Boris I. Shraiman*,†,1
*Kavli Institute for Theoretical Physics and †Department of Physics, University of California, Santa Barbara, California 93106

ABSTRACT Mounting evidence suggests that natural populations can harbor extensive fitness diversity with numerous genomic loci under selection. It is also known that genealogical trees for populations under selection are quantifiably different from those expected under neutral evolution and described statistically by Kingman's coalescent. While differences in the statistical structure of genealogies have long been used as a test for the presence of selection, the full extent of the information that they contain has not been exploited. Here we demonstrate that the shape of the reconstructed genealogical tree for a moderately large number of random genomic samples taken from a fitness diverse, but otherwise unstructured, asexual population can be used to predict the relative fitness of individuals within the sample. To achieve this we define a heuristic algorithm, which we test *in silico*, using simulations of a Wright–Fisher model for a realistic range of mutation rates and selection strength. Our inferred fitness ranking is based on a linear discriminator that identifies rapidly coalescing lineages in the reconstructed tree. Inferred fitness ranking correlates strongly with actual fitness, with a genome in the top 10% ranked being in the top 20% fittest with false discovery rate of 0.1–0.3, depending on the mutation/selection parameters. The ranking also enables us to predict the genotypes that future populations inherit from the present one. While the inference accuracy increases monotonically with sample size, samples of 200 nearly saturate the performance. We propose that our approach can be used for inferring relative fitness of genomes obtained in single-cell sequencing of tumors and in monitoring viral outbreaks.

MOST mutations are believed to have minimal effects on the fitness of the organism and much of the analysis of the genomic data on populations (see Excoffier and Heckel 2006 for a review of methods) has been based on the neutral hypothesis, according to which the dynamics of genetic polymorphisms and the overall genetic diversity of the population are governed by the neutral *drift*, *i.e.*, stochastic fluctuations in allele frequency arising from the intrinsic stochasticity in offspring number. The neutral model assumes that deleterious mutations are eliminated by selection fast enough to not significantly contribute to population diversity and beneficial mutations are rare enough to produce only occasional adaptive *sweeps*, where the population is taken over by the offspring of the adaptive genotype, transiently suppressing neutral genetic diversity. Statistical properties of genealogies generated by neutral dynamics in asexual populations are understood in great detail (Hein *et al.* 2005; Wakeley 2008) in terms of Kingman's *coalescent* process (Kingman 1982), which follows the ancestors of the present population back in time as far as the *most recent common ancestor* (MRCA). The neutral coalescent (Hein *et al.* 2005; Wakeley 2008) forms the basis for estimating mutation and recombination rates and provides the null hypothesis in tests for the presence of selection (Tajima 1989; Fu and Li 1993).

Yet, as advances in sequencing have made it possible to obtain quantitative data on genetic diversity, numerous studies have reached the conclusion that nonneutral polymorphisms are ubiquitous in populations across the spectrum of life: from viruses (Coffin *et al.* 1995; Novella *et al.* 1995; Moya *et al.* 2004; Neher and Leitner 2010; Batorsky *et al.* 2011) and bacteria (Barrick *et al.* 2009) to flies (Sella *et al.* 2009) to mitochondria (Seger *et al.* 2010) and cells in cancerous tumors (Merlo *et al.* 2006). In addition, laboratory evolution experiments in bacteria (Lenski *et al.* 1991; Miralles *et al.* 1999) and yeast (Kao and Sherlock 2008; Lang *et al.* 2011) have demonstrated directly that large asexual populations contain numerous subclones that are continuously generated by mutation and compete for fixation. Thus, large asexual populations cannot be assumed selectively neutral.

doi: 10.1534/genetics.113.160986
Manuscript received December 22, 2013; accepted for publication April 1, 2014; published Early Online April 26, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.160986/-/DC1.
[1]Corresponding author: Kavli Institute for Theoretical Physics, University of California, Santa Barbara, CA 93106. E-mail: shraiman@kitp.ucsb.edu

The presence of selection affects the shape of genealogical trees, often giving them an asymmetric and "comb-like" appearance that is strikingly different from that of the neutral trees generated by Kingman's coalescent (Hein *et al.* 2005; Wakeley 2008; Seger *et al.* 2010; Trevor *et al.* 2011). An example of such "genealogical anomalies"—*i.e.*, large deviations from neutral genealogical structure (Maia *et al.* 2004)—is provided by the recent study (Seger *et al.* 2010) of mitochondrial diversity in three distinct populations of whale lice, *Cyamus ovalis*, where the authors demonstrate that the observed genealogies are statistically consistent with a nonneutral model with frequent mutations of small selective effect.

Our analysis is based on a similar model of asexual evolutionary dynamics driven by small deleterious and beneficial mutations. In Figure 1 we show schematically a sample of continuous genealogy for a fixed-size population governed by Wright–Fisher dynamics (Hein *et al.* 2005; Wakeley 2008), incorporating genetic drift, mutation, and natural selection. The example in Figure 1 covers the period over which the offspring of one of the genomes (Figure 1, top) spread over the whole population (Figure 1, bottom). We ask, given a sample of genomes from the "present time" population (Figure 1, red circles), can one predict the genetic future of the population? Or, more specifically, can one identify, within the present sample, the closest relatives of the future population, *i.e.*, individuals that are on, or closest to, the genealogical backbone of the future population? Since long-term survival is correlated with fitness, this task is closely related to the problem of identifying the fitter fraction of the present-day sample.

Here, we demonstrate that the anomalous structure of the genealogical tree reconstructed for a sample of genomes can serve not only as the evidence of selection, but also as the basis for inferring the relative fitness ranking of sampled individuals and their proximity in sequence space to the fittest genomes. Information pertinent to this inference is contained in the pattern of coalescence for different lineages: in a nutshell, lineages that undergo several coalescence events much before others are relatively fit, while the less fit lineages do not merge with the rest (going backward in time) until later. Below we provide the simulation-based evidence supporting this scenario.

Our study builds upon considerable recent progress in the theoretical understanding of natural selection and drift dynamics in fitness-diverse asexual populations (Tsimring *et al.* 1996; Rouzine *et al.* 2003, 2008; Desai and Fisher 2007; O'Fallon *et al.* 2010; Sniegowski and Gerrish 2010; Good *et al.* 2012; Goyal *et al.* 2012; Walczak *et al.* 2012) and the emerging description of corresponding genealogies (Bolthausen and Sznitman 1998; Brunet *et al.* 2007; Berestycki 2009; O'Fallon *et al.* 2010; Seger *et al.* 2010; Desai *et al.* 2013; Walczak *et al.* 2012; Neher and Hallatschek 2013; Neher 2013). We focus on the asexual case and address how this approach might be extended to the analysis of recombining populations in the *Discussion*.

We focus on the regime where numerous beneficial or deleterious mutations segregate simultaneously and the population is formed by many clones with different fitness values.

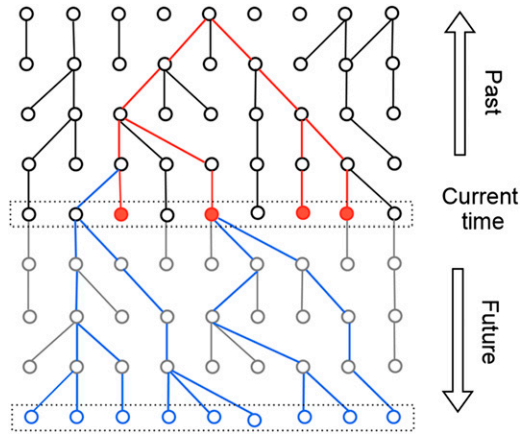

Figure 1 Schematic example of a genealogical trajectory, from past into the future, of an asexual population with fixed size ($N = 9$) and nonoverlapping generations. Nodes represent individual genomes, each linked to its ancestor in the previous generation. The example illustrates coalescence of the lineages of the bottom population toward its MRCA within the top population. The genealogical tree of a random sample (red) from the "current time" population partially overlaps the genealogy of the future population (blue). While actual ancestors of the future population (shown in blue) may or may not fall into the current sample, one can still define sample members that are closest to the surviving lineages. Identifying close relatives of future populations is the goal of our study.

In this regime, sometimes referred to as clonal interference (Miralles *et al.* 1999; Desai and Fisher 2007), competition between clones and the linkage between mutations play a key role in evolutionary dynamics. This regime is realized in large populations with high mutation rates. Precise conditions depend on the distribution of fitness effects of mutations and have been discussed in many recent articles (Rouzine *et al.* 2003, 2008; Desai and Fisher 2007; Brunet *et al.* 2008; Sniegowski and Gerrish 2010). For example, in the case where only beneficial mutations are present, the condition for being in the interference regime is given by $N\mu_b > 1/\log(Ns)$, where μ_b is the beneficial mutation rate (Desai and Fisher 2007; Brunet *et al.* 2008; Rouzine *et al.* 2008). This is basically the condition that new beneficial mutations get established in the population at a rate faster than they can "sweep" the population (see supporting information, File S1, for additional discussion). In the case of purifying selection where only deleterious mutations are present, it can be shown (Rouzine *et al.* 2003, 2008; Walczak *et al.* 2012) that the required condition is $N\exp(-\mu_d/s) < \frac{1}{s}\log(\mu_d/s)$, where μ_d is the deleterious mutation rate, s is the deleterious effect of mutations, and N is the population size.

Quite generally, when the population is formed by several clones with different fitness values, the fate of any new mutation depends not only on its own selective effect, but also on the fitness of the genotype on which it occurs (Good *et al.* 2012). As a result, the MRCA of such a fitness-diverse population is with high probability among the very fittest of its generation (O'Fallon *et al.* 2010). In return, the pattern of genealogical coalescence is controlled by the time it takes for surviving lineages to converge, as they are tracked back in time, on the leading edge of the fitness distribution at previous times.

This article is organized as follows. After formulating the model, we provide examples of genealogies, illustrating their anomalous shape compared to the neutral coalescent, and demonstrate the correlation between the ancestral *weight*, defined as the fraction of the present-day sample constituted by the descendants of the ancestor, and the mean fitness of the those descendants. We then define a fitness-ranking score based on the suitably integrated ancestral weights along the reconstructed lineage of each individual in the sample. Applying the ranking to numerous samples (for populations with the same and with different mutation/selection parameters) and comparing each realization to the true fitness known from the forward simulation, we demonstrate the ability of the proposed algorithm to infer the relative fitness of sampled genomes and to identify genotypes that are likely to survive into the future. The *Discussion* addresses possible applications and generalizations of the proposed inference method.

Model and Methods

Model of evolutionary dynamics

Consider an asexual population of size N that evolves with nonoverlapping generations under the influx of deleterious and beneficial mutations. New mutations arise at the rate $\mu + \mu_0$ (per genome per generation) with a fraction $\varepsilon\mu$ being beneficial, $(1 - \varepsilon)\mu$ deleterious, and the remainder μ_0 being neutral. For simplicity we assume both beneficial and deleterious mutations to have the same effect $s \ll 1$ and to change the fitness of individual i carrying that mutation additively: $F_i \to F_i \pm s$. As in the Wright–Fisher model, natural selection acts by biasing the probability of an individual genome to appear in the next generation, which is taken to be proportional to $\exp(f_i)$ with $f_i = F_i - \overline{F}$ being the individual fitness relative to the mean fitness of the population $\overline{F}$, which in general is a function of time.

We carried out 10^3 simulations of 2×10^5 generations for several plausible parameter combinations in the range of $\mu = 10^{-4}$–10^{-2}, $s = 10^{-3}$–10^{-2}, with ε taking values 0, 0.1, and 1, and $\mu_0 = 10\mu$ and $N = 64{,}000$. In File S1, we study the degree of clonal diversity and interference for the set of parameters that we have simulated and show that it explores a broad range in the clonal interference regime.

The genealogical trees were constructed in two ways. We recorded the genealogies in the course of the forward simulation, providing exact ancestries of any sample in the population. In addition, an inferred genealogy of random samples (between 30 and 500 genomes) was constructed using standard neighbor-joining/UPGMA-derived methods (Durbin 1998) is detailed in File S1. In File S1, we present the performance of the tree reconstruction method for different parameter values and show that it satisfactorily reconstructs the genealogical trees. For higher mutation rates (*e.g.*, $\mu = 5 \times 10^{-3}$ and $\mu = 10^{-2}$) where there are tens to hundreds of differences between a typical pair of genomes, even setting the neutral mutation rate equal to μ would be sufficient for an accurate reconstruction of the trees.

Fitness distribution and distortion in the shape of genealogical trees In the parameter range considered, simulated populations exhibit substantial fitness diversity with fitness variance in the order of $\sigma = \left(\frac{1}{N}\sum_{j=1}^{N} f_j^2\right)^{1/2} \approx 10^{-3}$–$10^{-2}$ arising from $\sim 10-10^3$ simultaneously segregating nonneutral polymorphisms. Figure 2, A and B, shows examples of the population-wide fitness distribution for two different mutation rates (see File S1 for additional examples). In general, genetic diversity in the population is an increasing function of μ/s. For the highest mutation rate and lowest selection coefficients considered, $\mu = 10^{-2}$ and $s = 10^{-3}$, the population exhibits extensive genetic diversity and is formed by many small clones (Figure 2B), whereas for the lower mutation rates, as in Figure 2A, the population typically includes larger clones.

Figure 2, D and E, shows typical examples of genealogical trees constructed for random samples of size $n = 30$ drawn from the populations corresponding to Figure 2, A and B, respectively. The fitness of sampled genomes, which we know from the forward simulation, is visualized using color. Also shown are ancestral weights along some of the lineages. This weight, w_i, is defined as the number of genomes in the present time sample that are direct descendants of lineage i. For example, each leaf at the bottom has weight $w = 1$, while the lineage at the root has the full weight of the sample $n = 30$. For the sake of comparison, we also show a typical genealogical tree for a neutrally evolving population in Figure 2F.

One immediately notes two well-known differences distinguishing Figure 2D and 2E from Figure 2F. Genealogies from fitness-diverse populations (i) have long terminal legs and are compressed toward the MRCA root of the tree and (ii) exhibit strong asymmetry of branching. These anomalies are quantified in Figure 3. Figure 3A presents distributions of pairwise coalescent times in the population, τ_{ij}, for $\{i, j\}$ genome pairs for several parameter sets. In Kingman's coalescent, τ_{ij} has an exponential distribution (with mean N) (Hein *et al.* 2005; Wakeley 2008) and most lineages in a genealogical tree coalesce at early times (looking backward). In contrast, the bulk of coalescence in a population under selection is significantly delayed compared to the total coalescent time—an effect corresponding to the comb-like appearance of the trees.

The asymmetry of branching is quantified in Figure 3B, which presents the distribution of weights at the level just below the root, where there are only two ancestral lineages left in the tree. The strong bias toward extreme values of w in populations under selection is contrasted with w-independent distribution predicted and observed in the neutral case (see File S1 for additional characteristics that quantify differences between the shapes of trees).

Results

Correlation between ancestral weight and offspring fitness

Let us consider the whole population and trace the surviving lineages back in time, identifying all ancestors of the present-day

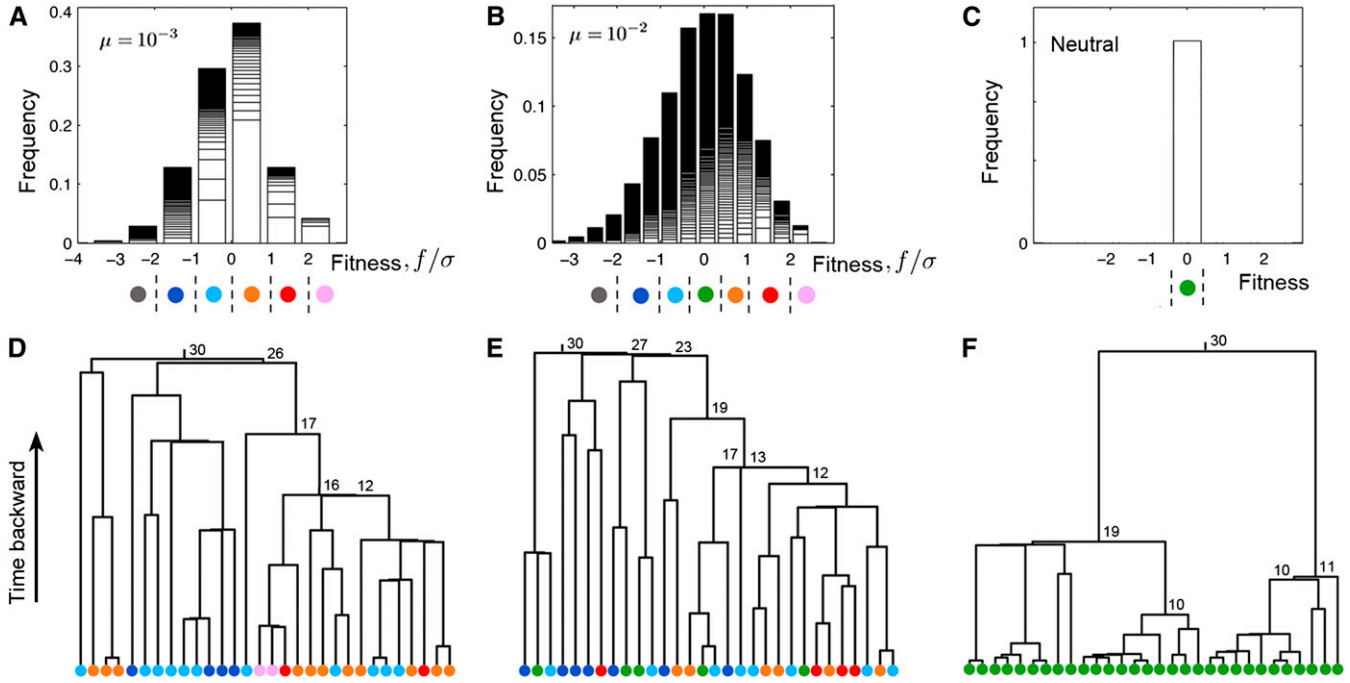

Figure 2 Fitness distributions and examples of genealogical trees. (A) Fitness distribution at one time point for a population with $\mu = 10^{-3}$, $s = 2 \times 10^{-3}$, and $\sigma \simeq 2.2 \times 10^{-3}$. Each bin corresponds to a fitness class and each class is composed of multiple clones delineated by horizontal lines within each bar, with larger clones stacked on the bottom. (Here, clones are defined using only the nonneutral mutations.) Also shown is the color code used in D. (B) Same as A but for a higher mutation rate $\mu = 10^{-2}$ and $\sigma \simeq 5 \times 10^{-3}$. (C) Same as A but for a neutral population. (D) A typical genealogical tree for a random sample of size $n = 30$ from the same population as in A. Each circle corresponds to one sampled genome and the color represents its fitness. Branch lengths are drawn in linear proportion to the corresponding time interval. Numbers next to internal nodes are the weights of the corresponding ancestors (only weights >10 are shown). Note the striking asymmetry of branching, *i.e.*, uneven distribution of weight among the two lineages descending from each internal node. (E) Same as D but for the population shown in B. Note that the colors (gray and plum) corresponding to the extremes of the distribution (B) are absent from the small sample shown. (F) Same as D but for a neutral population. To focus on the shape of the genealogy, we have normalized the "height" of the trees in D–F to the time to the MRCA, which makes the neutral tree appear as tall as the trees for the populations under selection (whereas the coalescence time is really much longer in the neutral case). Note the short terminal legs and more symmetric branching. $N = 64{,}000$ and $\varepsilon = 0.1$ for A–F.

population t generations in the past. Figure 4A shows the distribution of the ancestral fitness (relative to the mean for that generation) at several time points in the past. This distribution becomes progressively shifted toward higher fitness values compared to the distribution for the whole population (O'Fallon *et al.* 2010). In the limit of large times, this distribution converges to the nonextinction probability as a function of the fitness in the ancestral population (Neher *et al.* 2010; O'Fallon *et al.* 2010; Neher and Hallatschek 2013).

Let us consider the time in the past when there are still a large number of ancestors (*e.g.*, $\sim 10^3$ in the population of $N = 64{,}000$, which under conditions corresponding to simulations in Figure 4A occurs at $t \simeq 100$). Figure 4B shows the scatter plot of the weight of ancestors *vs.* their fitness advantage. Note that, by collapsing the points on the fitness axis, one gets the histogram shown in Figure 4A. We observe a strong positive correlation between the weight and the fitness of an ancestor. Higher-fitness individuals in the past generations are not only more likely to survive, but, conditioned on survival, they also leave more offspring. Thus the weight of the ancestor, which can be determined from a reconstructed genealogical tree, can be used as a proxy for ancestral fitness: a quantity that one does not expect to know directly, except in the case of computer simulations. In File S1 we provide plots of average ancestral fitness conditioned on its weight for various time points and parameter sets and confirm that the positive correlation between the weight and the fitness of ancestors holds quite generally. This correlation decreases as the time shifts farther into the past.

Next, we examine the correlation between the weight of an ancestor and the fitness of its surviving progeny. Consider a sample of genomes with size n and the corresponding genealogical tree. One expects genomes that are derived from relatively high-fitness ancestors to belong to higher-fitness classes at the present time. Since ancestral fitness correlates with weight, we expect higher-weight ancestors to produce, on average, higher-fitness descendants. To see this, let us consider an ancestor i, with weight w_i, that existed some t generations in the past. We examine the fitness $\{f_1, \ldots, f_{w_i}\}$ of the w_i offspring in the sample descending from that ancestor. In particular, we focus on the mean, $F(w_i) = (1/w_i)\sum_{j=1}^{w_i} f_j$, and the variance, $\Sigma^2(w_i) = (1/w_i)\sum_{j=1}^{w_i}\left(f_j - F(w_i)\right)^2$, over the w_i offspring (subscript d refers to descendants). Let us denote the

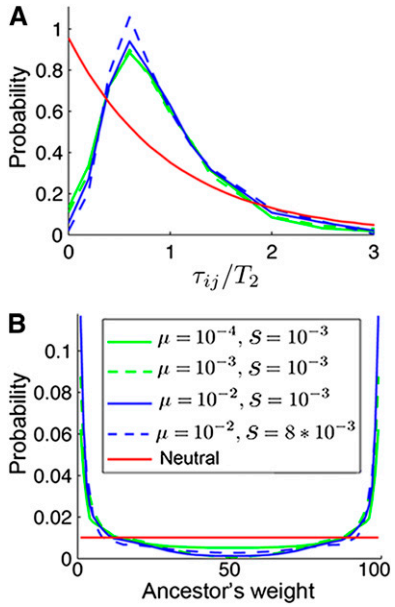

Figure 3 Distortion in the shape of genealogies in the presence of selection. (A) Distribution of pairwise coalescent time, scaled with its mean, T_2. (B) Probability of an ancestor to have weight w when there are $a = 2$ lineages left in the genealogical tree of $n = 100$ samples. Distributions are based on 8000 random samples and population replicas. $N = 64{,}000$ and $\varepsilon = 0.1$ in both A and B.

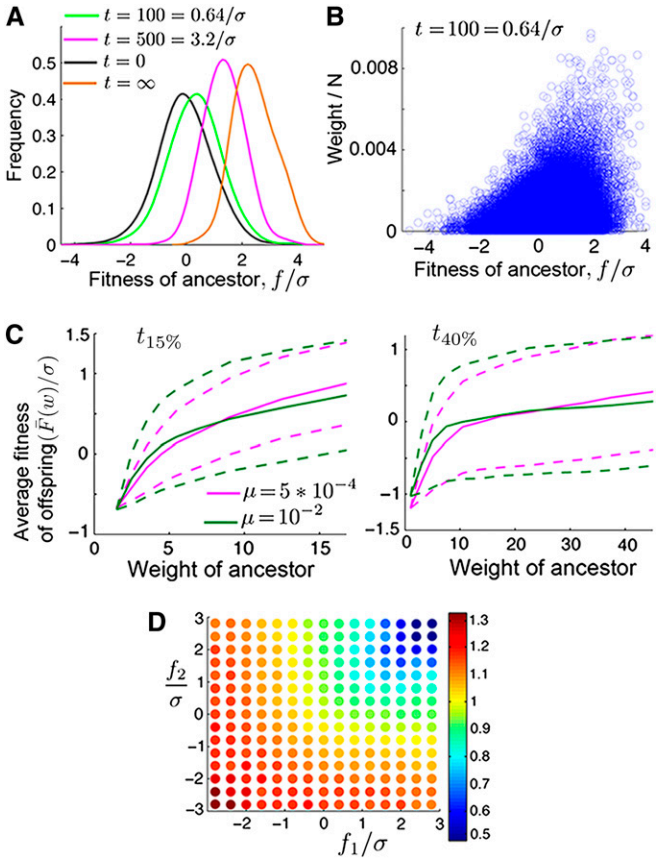

Figure 4 Correlation between fitness, weight, and coalescent time. (A) Fitness distribution of the ancestors of the whole population, for several time points in the past. (B) Scatter plot of weight *vs.* ancestral fitness $t = 100$ generations back. (C) Solid lines show average fitness of offspring as a function of the ancestral weight in a sample of size $n = 100$ at two different time slices in the past. Dashed lines represent the standard deviation, $\overline{\Sigma}(w)/\sigma$, above and below the mean, $\bar{F}(w)/\sigma$. The time slices ($t_{15\%}$ and $t_{40\%}$) were chosen to be the first time (looking backward) when genealogy contained a lineage with weight >15% or 40% of n, respectively. (D) Heat map of mean pairwise coalescent time as a function of the fitness of the involved genomes, $f_{1,2}$, normalized by the mean pairwise coalescent time for the whole population: $t_2(f_1, f_2)/T_2$. Evolutionary model parameters are $N = 64{,}000$, $\varepsilon = 0.1$, and $s = 2 \times 10^{-3}$ in A–D; $\mu = 10^{-3}$ in A, B, and D.

average of these quantities over random samples of genomes and over population replicas by $\overline{F}(w_i) = \langle F(w_i) \rangle$ and $\overline{\Sigma}^2(w_i) = \left\langle \Sigma^2(w_i) \right\rangle$. Note that $\overline{F}(w)$ and $\overline{\Sigma}(w)$ depend on the time t, namely, how far back in the genealogy one is considering.

In Figure 4, C and D, we show $\overline{F}(w)/\sigma$ and $\overline{\Sigma}(w)/\sigma$ at two different time points in the past for samples of size $n = 100$ (see File S1 for other parameter sets). In both cases, the mean fitness of the derived genomes is an increasing function of the weight of their ancestor. Consider a time close to the root of a tree such that a lineage can have a weight that is a significant portion of the sample size (*e.g.*, right plot in Figure 4C). As expected, the value of $\overline{F}(w)$ for such high-weight ancestors is close to zero (remember that f_i was defined relative to the population mean, so that the average of f_i over the whole sample is zero). At the same time $\overline{\Sigma}(w)/\sigma \to 1$ for ancestors with w approaching n. Interestingly, for the lineages that still have a small weight late in the coalescence process, the value of $\overline{F}(w)$ is clearly negative.

High-fitness genomes typically merge first in a tree and form high-weight ancestors. To make this point clear, consider the distribution of the pairwise coalescent time, τ_{ij}, shown in Figure 3A. Averaging τ_{ij} over all $\{i, j\}$ pairs of genomes in a population gives the mean coalescent time T_2. Now, consider the average of τ_{ij} conditioned on the fitness of the two genomes and denote it by $t_2(f_i, f_j)$. Figure 4D shows a heat map of $t_2(f_i, f_j)/T_2$. For two genomes both with high fitness, the average coalescent time is $< T_2$. The reason is that such genomes are likely to be relatively recent lineages emanating from the "nose" of the distribution. In other words, the chance of sampling identical or similar sequences is greater for fitter samples than for less fit samples, since fitter samples have shorter average pairwise coalescent time. This observation is the key to the proposed fitness inference method.

Relative fitness inference based on the reconstructed genealogy

Above we have reviewed different ways in which the shape of the genealogical trees for populations under selection differs from the expectation of neutral theory. We have also demonstrated the correlation between ancestral weights and the fitness of the descendants. We showed that sampled genomes that belong to high-fitness classes typically have shorter coalescent time compared to unfit genomes. We now show that this insight can be converted into a method for inferring relative fitness of genomes within the sample.

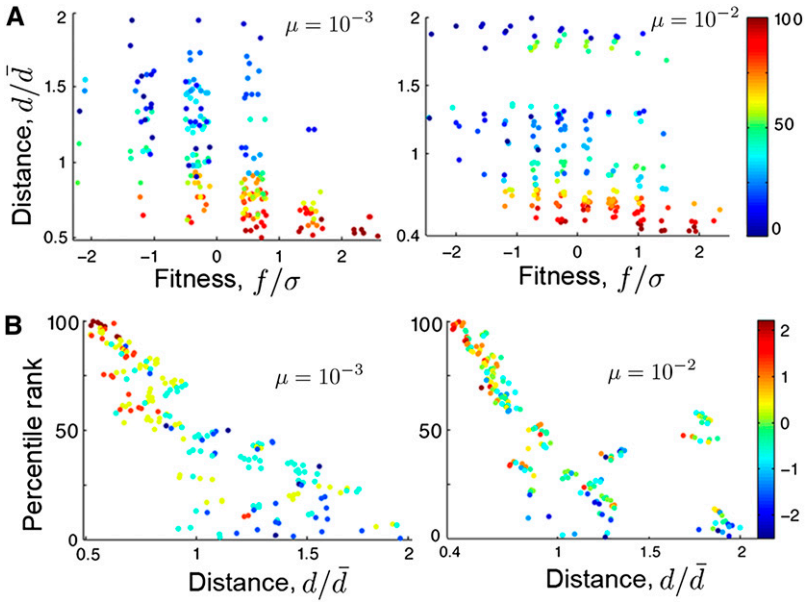

Figure 5 Examples of performance of the ranking algorithm. (A) Heat map of rank as a function of fitness and average distance to the top 10% of fittest genomes. Distance d is normalized by its mean $\bar{d}$. Left and right panels correspond to two samples of size $n = 200$ drawn from the same populations as in Figure 2A ($\mu = 10^{-3}$) and Figure 2B ($\mu = 10^{-2}$), respectively. To avoid overlap of points, a small random number has been added to the fitness coordinate of each point. (B) Scatter plot of rank *vs.* distance to the top 10% of fittest genomes (color map represents f/σ). The plots correspond to the same trees as in A. $N = 64{,}000$ and $\varepsilon = 0.1$ in A and B.

To that end, let us consider a randomly chosen set of n genomes from a population and use standard phylogenetic tree-building methods (see File S1) to approximately reconstruct the genealogy of the sample. The accuracy of the reconstructed genealogy compared to the actual genealogy, known exactly from the forward simulation of population dynamics, is discussed in File S1. It increases with the neutral mutation rate μ_0: in the biologically plausible regime of $\mu_0/\mu \approx 10$ considered here, it proves more than adequate to enable meaningful inference.

Next, based on the reconstructed tree, we associate with each leaf $i = 1, \ldots, n$ a *fitness-proxy score* (FPS), ϕ_i, defined by its lineage within the tree. Specifically, we define ϕ_i as a linear discriminator in the form

$$\phi_i = \sum_{k=1}^{m_i} \Theta\left(\frac{t_{a_k(i)}}{T_2}\right)\left[w_{a_k(i)} - w_{a_{k-1}(i)}\right], \qquad (1)$$

where $\{a_k(i)\}$ is the lineage of genome i, starting with the genome itself as $a_0(i)$ and running the length, m_i, of the lineage (*i.e.*, the number of nodes) until the root of the tree. When the ancestral lineage $a_{k-1}(i)$ with weight $w_{a_{k-1}(i)}$ coalesces at internal node k, it forms a new ancestral lineage $a_k(i)$ with weight $w_{a_k(i)}$ (see File S1 for an illustration of this notation on the example of a particular tree). The time of formation of the corresponding internal node is denoted by $t_{a_k(i)}$. The parameter T_2 is the estimate of the average pairwise coalescent time, obtained from the sampled genomes. Finally, $\Theta(x)$ is a "soft step" function (a.k.a. Fermi function): $\Theta(x) = (1 + \exp(\beta(x/x_* - 1)))^{-1}$ parameterized by the position of the step x_* and its characteristic width β. If $\beta \gg 1$, the function $\Theta(x)$ changes abruptly from one to zero as x becomes $> x_*$, so that $\phi_i = w_{a_*} - 1$, where a_* is the oldest ancestor in the lineage with $t_{a_*} < x_* T_2$. For $\beta \sim 1$ the FPS is defined by a weighted sum of ancestral weights (see File S1 for details).

The logic behind our heuristic choice of the specific form of ϕ_i is to exploit the correlation between the offspring fitness and ancestral weights. Note that, at least on the high-fitness/high-weight end of the distribution, this correlation decreases as t_a becomes large compared to T_2. The reason for this is that for long times in the past, even the lineages originating from high-fitness ancestors spread all over the fitness distribution at the present time. Hence, we choose $x_* < 1$: specifically the results below were obtained with $x_* = 0.5$ and $\beta = 5$, but in File S1 we examine the performance of the ranking algorithm as a function of the parameters and demonstrate that nearly optimal performance for the present form of the FPS is achieved for a broad range of x_* and β. Critically, normalization of t_a to the characteristic time of coalescence for the sample, T_2, eliminates the need to know the evolutionary parameters of the population, such as μ or N.

We rank genomes according to their ϕ_i score and compare this ranking with the actual fitness of each genome. In addition to inferring relative fitness, it is useful to know how genetically close a genome with a given rank is to the fittest one in the sample. Hence, for each genome we define d_i as the average of its Hamming distance to the fittest 10% of genomes in the sample. Figure 5, A and B, shows the results of the ranking for two $n = 200$ samples from the populations that already appeared in Figure 2, A and B. We observe a correlation between FPS ranking and the actual fitness in general and the tendency (quantified below) for the fittest genomes of the sample to show in the top ranks. In addition, high-ranked genomes that do not belong to high-fitness classes still have small d_i values, indicating that they are genetically close to the fittest genotypes. In other words, even if a high-ranked genome is not fit, typically it has only recently branched off from a fit clone and, compared to a randomly chosen unfit sequence, shares greater sequence similarity with fittest genotypes.

The above observations are confirmed and quantified by repeating and averaging the analysis for 8000 independent population samples and different sets of parameters. Specifically, Figure 6 shows the fitness distribution of the top-ranked genomes for the two parameter sets used in Figure 5. The results clearly indicate that the top-ranked genomes tend to be among fitter genotypes in the population. In addition, Figure 7A shows mean fitness conditional on the FPS ranking and Figure 7B shows the mean rank conditional on actual fitness (normalized by σ) for two different values of μ. Figure 7C shows mean distance from the fittest conditional on the FPS ranking (for four different values of μ), with distance normalized to $\Delta_{10\%}$ defined as the average d_i among the fittest 10%. Remarkably, we observe that $d/\Delta_{10\%}$ for the highest-ranked genomes approaches one, indicating good convergence, in the sense of Hamming distance, of the top-ranked genomes to the fittest set. Further analysis of the algorithm's performance, as well as additional parameter sets including the case of purifying selection ($\varepsilon = 0$), can be found in File S1.

As already mentioned, we are interested in the set of evolutionary parameters for which many mutations segregate simultaneously and the population is formed by numerous clones with different fitness values. The opposing limit, which occurs for small population size, N, or mutation rate, μ, corresponds to the regime of selective sweeps/successive mutations. In this latter regime, the population is typically formed by only a few clones and the fitness diversity is relatively low. Moreover, for smaller values of the parameters N and μ, the inference of genealogical trees becomes less accurate as the genetic diversity between sampled sequences decreases. Therefore, we expect the performance of the algorithm to deteriorate for small population size, N, or mutation rate, μ. In File S1, we show that for smaller values of the quantity $\theta = N\mu$, particularly for $\theta < 1$, the performance of the fitness inference algorithm deteriorates. We also show that the performance of the algorithm deteriorates as the fitness diversity in the population, represented by σ/s, decreases. Note that the quantity σ/s provides a measure for the number of different fitness classes in the population.

As we see in Figure 5A, high-ranked genomes that do not belong to fittest classes still tend to have small genetic distance to fittest individuals (also note in Figure 2, D and E, the genomes with blue color located close to the mostly orange/red clusters on the right side of the trees). This is because the Hamming distance is dominated by neutral mutations $\mu_0 \gg \mu$ and is less susceptible to fluctuations compared to fitness, which is defined by a much smaller number of nonneutral mutations. To the extent that genetic relatedness is defined by the distance, the latter is essential for identifying within the sample the closest relatives of future populations. Taking advantage of ready accessibility of evolutionary future within our simulations, we have directly tested the ability of our approach to identify, within the sample, the genotypes that are closer to those of future

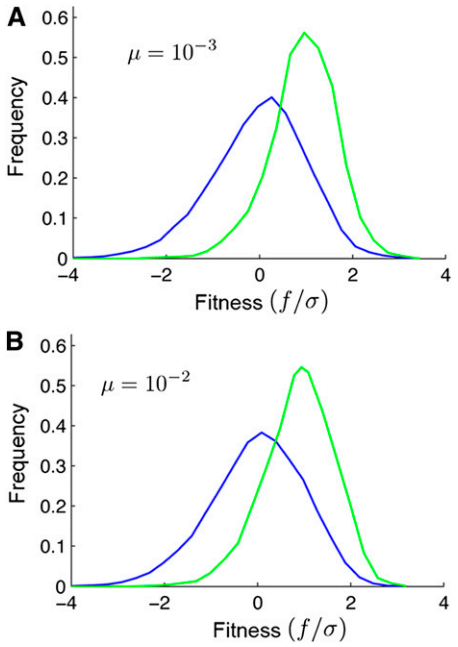

Figure 6 Fitness distribution of the top 10% ranked genomes (green) compared to the fitness distribution of the whole population (blue). A and B correspond, respectively, to $\mu = 10^{-3}$ and $\mu = 10^{-2}$; other population parameters are the same as in Figure 5. The distributions are obtained by averaging over 8000 population replicas.

populations. For each sampled genome, we define d_i' as the average of its Hamming distance to all of the genomes in the current population that are ancestors of the population in a generation about one genetic turnover time in the future (we know these ancestors from the forward simulation). We choose this turnover time to be the first time in the future when <1% of individuals from the current population have any descendant left. In each case we normalized the distances by $\Delta'_{10\%}$, defined as the average of the smallest 10% of values of d_i'. Figure 7D shows $d'/\Delta'_{10\%}$ conditional on the FPS ranking. We again observe that $d'/\Delta'_{10\%}$ for the highest-ranked genomes gets close to one, indicating that the top-ranked genomes are indeed close to the ancestors of future generations. This means that the FPS ranking makes it possible to identify the genetic elements (common among the high-rank genomes) that future populations inherit from the present one.

Finally, we examine the fitness of the genomes with the 10% highest rank. Consider the sorted vector $F = [f_1, \ldots, f_n]$ that contains the actual fitness values for all the sampled genomes. In Figure 7E, we show that the probability for the fitness of a genome within the top 10% rank to be above the median fitness is ~0.9, for the broad range of parameters considered. The probability for the fitness of a top 10% -ranked genome to belong to the top 20% fitness class is given by the solid lines in Figure 7F and is >0.7. Note that some of the sampled genomes can have equal fitness (*i.e.*, F contains duplicate values), which is more common for lower mutation rates where the fitness diversity in the populations is limited. Hence, to provide a meaningful comparison for this probability,

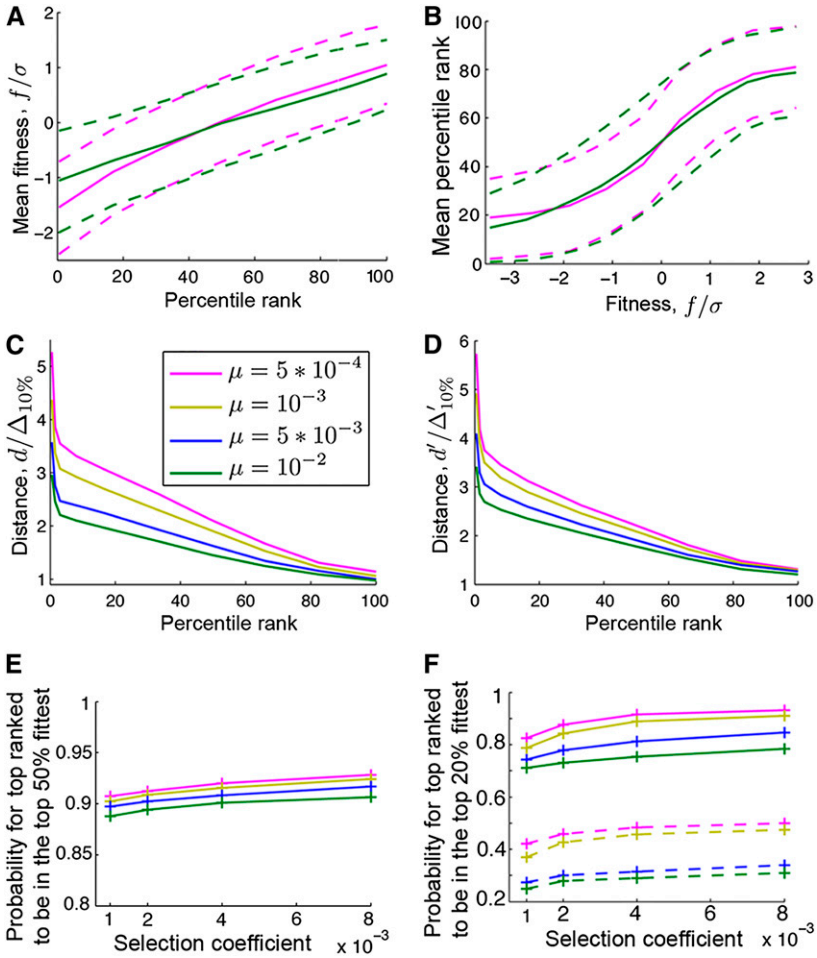

Figure 7 Performance of the fitness-ranking algorithm. (A) Solid lines show mean fitness as a function of rank. Dashed lines show standard deviation above and below the mean ($\mu = 5 \times 10^{-4}$ and 10^{-2}; see inset in C). (B) Same as A for mean rank as a function of fitness. (C) Mean Hamming distance to the top 10% fitness set, normalized by $\Delta_{10\%}$ (see text) as a function of rank. (D) Mean Hamming distance to ancestors of the generation at one turnover time in the future, normalized by $\Delta'_{10\%}$ (see main text) as a function of rank. (E) Probability for the fitness of a genome within the top 10% ranked to belong to the top 50% of fitness values of sampled genomes for a range of mutation rates and selection coefficients. (F) Probability for the fitness of a genome within the top 10% ranked to belong to the top 20% of fitness values shown using solid lines. The dashed lines show this probability for a randomly chosen genome (see main text). Sample size $n = 200$, $N = 64{,}000$, and $\varepsilon = 0.1$ in all cases; $s = 2 \times 10^{-3}$ in A–D.

in Figure 7F we show—using dashed lines—the probability for a random genome to be in the top 20% fittest.

In summary, the above results clearly indicate the power of the proposed inference method. The performance of the method improves monotonically with increasing sample size (see File S1): it degrades significantly, compared to the results presented above, for $n < 100$ but approaches saturation for $n > 200$.

As we discussed earlier, we are interested in the set of evolutionary parameters for which several mutations segregate simultaneously and the population is formed by several clones with varying fitness values. In the opposing limit corresponding to the regime of selective sweeps/successive mutations, we expect the performance of the algorithm to deteriorate, as some fundamental aspects of the dynamics (such as the dependence of the fate of mutations on the genetic background) are different. To make this point clear, we calculated the Pearson correlation coefficient between the rank and the distance d'. Figure 8A shows this correlation as a function of the parameter $N\mu$. As we see, for smaller values of $N\mu$, particularly for $N\mu < 1$, the correlation coefficient drops significantly.

Similarly, in Figure 8B, we show the above correlation as a function of the participation fraction, defined as the probability that two randomly chosen genomes belong to the same clone [*i.e.*, $\left\langle \sum_{i=1}^{c} (n_i/N)^2 \right\rangle$, where n_i is the size of the ith clone]. Note that the participation fraction gives a measure for the fitness diversity in the population. Figure 8B shows again that as the genetic diversity in the population decreases, the correlation coefficient between the rank and the distance d' drops as well.

Discussion

Whereas one often thinks of evolution occurring on geological timescales, evolutionary dynamics can also unfold swiftly as they do in bacteria acquiring antibiotic resistance, in human immunodeficiency virus evading Cytotoxic T-Cell response in the course of infection, or in the progression of an aggressive cancer. Recent advances in sequencing (Smith *et al.* 2010; Navin *et al.* 2011) have made it possible to extensively sample such rapidly evolving populations. The amount and quality of genomic data on populations will only continue to increase, accentuating the challenge of extracting more information from sampled genomes. Here, we have demonstrated that the shape of genealogical trees contains much more information than merely the evidence for (or against) selection within a population. As a proof-of-principle we have formulated a method for ranking the relative fitness of individual

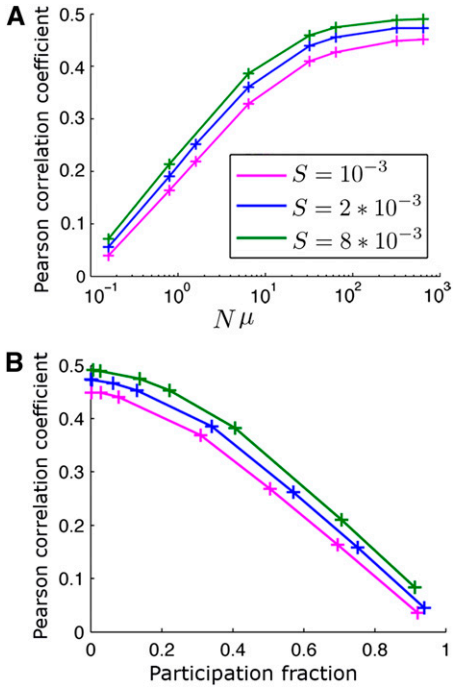

Figure 8 As the fitness diversity in the population decreases, the performance of the fitness-ranking algorithm deteriorates. (A) Correlation between the rank and the distance d' (Hamming distance to the ancestors of the future generations in the current population) as a function of $N \times \mu$. (B) Same as A but the *x*-axis represents the participation fraction, defined as the probability that two randomly chosen genomes belong to the same clone. $\varepsilon = 0.1$.

genomes sampled from a fitness-diverse but otherwise unstructured population, in the absence of any information other than genomic sequence. This provides the possibility of forecasting the common genotype of the future on the timescale of genetic turnover.

Our demonstration was based on a vast simplification of biological and ecological reality. Our model assumed fixed population size and constant environment; it neglected epistasis and assumed all nonneutral mutations (both deleterious and beneficial) to have the same selective strength. While we have explored a biologically interesting range of parameters within the considered model, it would be useful to extend the study to a broader class of models. Yet, we expect the proposed method to be quite robust, because it is based on the very fundamental aspect of evolutionary dynamics, realized when the population size and the mutation rate are sufficiently large. Under such a condition, the population harbors substantial nonneutral diversity, and fitness differentials between individuals are formed by the contributions of numerous weakly selected loci rather than a small number of strong ones. In this multilocus weak selection regime, surviving lineages in the course of time move from the nose of the fitness distribution toward the center, in a biased diffusion fashion. The correlation between early coalescence and rapid increase of ancestral weight along the lineages with high relative fitness derives from the continuous genetic turnover of the population described above. This turnover occurs in traveling-waves models corresponding to the continuous adaptation scenario (Tsimring *et al.* 1996; Rouzine *et al.* 2003), in the dynamic mutation–selection balance (Goyal *et al.* 2012) that involves both deleterious and compensating beneficial mutations, and in the case of purifying selection ($\varepsilon = 0$) (Gordo and Charlesworth 2000; Rouzine *et al.* 2003, 2008; Walczak *et al.* 2012).

A detailed statistical analysis of the way lineages propagate along the fitness axis could allow us to improve FPS by optimizing the trade-off between gaining more information about a particular lineage by tracking it farther back in time and the loss of predictive power due to the fact that beyond the genetic turnover time even lineages of the fittest ancestors spread all over the fitness distribution. Presently we have dealt with the problem heuristically by focusing on the coalescence sequence for each lineage up to $\sim 0.5T_2$. The advantage of our simple heuristic approach is that it is more likely to be model independent than the more fine-tuned methods. Building on the recent progress in understanding of genealogies in the presence of multilocus selection (O'Fallon *et al.* 2010; Walczak *et al.* 2012; Neher and Hallatschek 2013), it should be possible to replace our heuristic approach by a more systematic one.

It would be interesting to extend the fitness inference method to recombining populations. This should be relatively straightforward as long as genetic turnover time is fast compared to the inverse recombination rate. For a chromosome with an approximately uniform crossover probability, this condition defines a characteristic length below which loci coalesce in essentially recombination-free genealogies (Neher *et al.* 2013). Roughly, the asexual coalescent considerations would apply to a 1-cM size locus provided that it harbors $\sigma > 10^{-2}$. More careful analysis is, however, necessary to deal with the Hill–Robertson effect or *genetic draft* (Hill and Robertson 1966; Neher and Shraiman 2011) caused by the transient linkage of the locus to the rest of the genome, which effectively adds noise, reducing effectiveness of selection on the individual loci.

The highest priority for the future would be to test the method on experimental or epidemiological data. Applications are possible wherever genomic data are available for fitness-diverse, but otherwise unstructured populations. Genomic data from single-cell sequencing of tumors (Navin *et al.* 2011) or from localized influenza outbreaks (Squires *et al.* 2011) are among the interesting possibilities to be considered. For example, it would be interesting to compare the proposed method with the clustering-based approach of Plotkin *et al.* (2002) to predict antigenic evolution of influenza A. A challenge in applying our approach to the existing influenza virus data is posed by the possibility of strong geographical/temporal biases in the sampling patterns. In addition, there is a bias due to preferential sequencing of antigenically distinct genomes on the basis of HI assays (Bush *et al.* 2001). For certain analyses, such as measuring the average substitution rate over a long period, such biases are less important (Russell *et al.* 2008; Bhatt *et al.* 2011; Strelkowa and Lässig

2012), but a more principled method of addressing the sampling bias may be necessary to achieve the full potential of our method of fitness inference.

In addition to predicting which genotypes are more likely to appear in future generations, the fitness inference method could be used for QTL mapping (Broman and Sen 2009) with FPS-based ranking being the quantitative phenotype that could be used to identify highly adaptive or deleterious alleles.

Acknowledgments

We thank Richard Neher, Daniel Balick, and Sidhartha Goyal for many useful discussions. A.D. was supported by HFSP grant RFG0045/2010 and National Science Foundation grant PHY11-25915 while B.I.S. acknowledges support from National Institute of General Medical Sciences grant R01 GM086793.

Literature Cited

Barrick, J., D. Yu, S. Yoon, H. Jeong, T. Oh *et al.*, 2009 Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature 461: 1243–1247.

Batorsky, R., M. F. Kearney, S. E. Palmer, F. Maldarelli, I. M. Rouzine *et al.*, 2011 Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. Proc. Natl. Acad. Sci. USA 108: 5661–5666.

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

Berestycki, N., 2009 Recent progress in coalescent theory. Ensaios Matematicos 16: 1–193.

Bhatt, S., E. Holmes, and O. Pybus, 2011 The genomic rate of molecular adaptation of the human influenza a virus. Mol. Biol. Evol. 28: 2443–2451.

Bolthausen, E., and A. Sznitman, 1998 On Ruelle's probability cascades and an abstract cavity method. Commun. Math. Phys. 197: 247–276.

Broman, K., and S. Sen, 2009 *A Guide to QTL Mapping with R/QTL (Statistics for Biology and Health)*. Springer-Verlag, New York.

Brunet, E., B. Derrida, A. Mueller, and S. Munier, 2007 Effect of selection on ancestry: an exactly soluble case and its phenomenological generalization. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. 76: 041104.

Brunet, É., I. M. Rouzine, and C. O. Wilke, 2008 The stochastic edge in adaptive evolution. Genetics 179: 603–620.

Bush, R. M., 2001 Predicting adaptive evolution. Nat. Rev. Genet. 2: 387–392.

Coffin, J. M., 1995 HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science 267: 483–489.

Desai, M., and D. Fisher, 2007 Beneficial mutation–selection balance and the effect of linkage on positive selection. Genetics 176: 1759–1798.

Desai, M., A. Walczak, and D. Fisher, 2013 Genetic diversity and the structure of genealogies in rapidly adapting populations. Genetics 193: 565–85.

Durbin, R., 1998 *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge/London/New York.

Excoffier, L., and G. Heckel, 2006 Computer programs for population genetics data analysis: a survival guide. Nat. Rev. Genet. 7: 745–758.

Fu, Y., and W. Li, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693–709.

Good, B., I. Rouzine, D. Balick, O. Hallatschek, and M. Desai, 2012 Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. Proc. Natl. Acad. Sci. USA 109: 4950–5.

Gordo, I., and B. Charlesworth, 2000 The degeneration of asexual haploid populations and the speed of Muller's ratchet. Genetics 154: 1379–1387.

Goyal, S., D. J. Balick, E. R. Jerison, R. A. Neher, B. I. Shraiman *et al.*, 2012 Dynamic mutation–selection balance as an evolutionary attractor. Genetics 191: 1309–1319.

Hein, J., M. Schierup, and C. Wiuf, 2005 *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, New York.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. 8: 269–294.

Kao, K., and G. Sherlock, 2008 Molecular characterization of clonal interference during adaptive evolution in asexual populations of Saccharomyces cerevisiae. Nat. Genet. 40: 1499–1504.

Kingman, J., 1982 The coalescent. Stoch. Proc. Appl. 13: 235–248.

Lang, G., D. Botstein, and M. Desai, 2011 Genetic variation and the fate of beneficial mutations in asexual populations. Genetics 188: 647–661.

Lenski, R., M. Rose, S. Simpson, and S. Tadler, 1991 Long-term experimental evolution in Escherichia coli. i. adaptation and divergence during 2,000 generations. Am. Nat. 138: 1315–1341.

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

Merlo, L., J. Pepper, B. Reid, and C. Maley, 2006 Cancer as an evolutionary and ecological process. Nat. Rev. Cancer 6: 924–935.

Miralles, R., P. Gerrish, A. Moya, and S. Elena, 1999 Clonal interference and the evolution of RNA viruses. Science 285: 1745–1747.

Moya, A., E. Holmes, and F. González-Candelas, 2004 The population genetics and evolutionary epidemiology of RNA viruses. Nat. Rev. Microbiol. 2: 279–288.

Navin, N., J. Kendall, J. Troge, P. Andrews, L. Rodgers *et al.*, 2011 Tumour evolution inferred by single-cell sequencing. Nature 472: 90–94.

Neher, R., and T. Leitner, 2010 Recombination rate and selection strength in HIV intra-patient evolution. PLoS Comput. Biol. 6: e1000660.

Neher, R., and B. Shraiman, 2011 Genetic draft and quasineutrality in large facultatively sexual populations. Genetics 188: 975–996.

Neher, R., B. Shraiman, and D. Fisher, 2010 Rate of adaptation in large sexual populations. Genetics 184: 467–481.

Neher, R. A., 2013 Genetic draft, selective interference, and population genetics of rapid adaptation. arXiv: 1302.1148.

Neher, R. A., and O. Hallatschek, 2013 Genealogies of rapidly adapting populations. Proc. Natl. Acad. Sci. USA 110: 437–442.

Neher, R. A., T. A. Kessinger, and B. I. Shraiman, 2013 Coalescence and genetic diversity in sexual populations under selection. Proc. Natl. Acad. Sci. USA 110: 15836–15841.

Novella, I. S., D. K. Clarke, J. Quer, E. A. Duarte, C. H. Lee *et al.*, 1995 Extreme fitness differences in mammalian and insect hosts after continuous replication of vesicular stomatitis virus in sandfly cells. J. Virol. 69: 6805–6809.

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

Plotkin, J., J. Dushoff, and S. Levin, 2002 Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. Proc. Natl. Acad. Sci. USA 99: 6263–6268.

Rouzine, I., J. Wakeley, and J. Coffin, 2003 The solitary wave of asexual evolution. Proc. Natl. Acad. Sci. USA 100: 587–592.

Rouzine, I., É. Brunet, and C. Wilke, 2008 The traveling-wave approach to asexual evolution: Muller's ratchet and speed of

adaptation. Theor. Popul. Biol. 73: 24–46.Russell, C., T. Jones, I. Barr, N. Cox, R. Garten *et al.*, 2008 The global circulation of seasonal influenza a (h3n2) viruses. Science 320: 340–346.

Seger, J., W. Smith, J. Perry, J. Hunn, Z. Kaliszewska *et al.*, 2010 Gene genealogies strongly distorted by weakly interfering mutations in constant environments. Genetics 184: 529–545.

Sella, G., D. Petrov, M. Przeworski, and P. Andolfatto, 2009 Pervasive natural selection in the Drosophila genome? PLoS Genet. 5: e1000495.

Smith, A., L. Heisler, R. Onge, E. Farias-Hesson, I. Wallace *et al.*, 2010 Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Res. 38: e142.

Sniegowski, P., and P. Gerrish, 2010 Beneficial mutations and the dynamics of adaptation in asexual populations. Philos. Trans. R. Soc. Lond. B Biol. Sci. 365: 1255–1263.

Squires, R., J. Noronha, V. Hunt, A. García-Sastre, C. Macken *et al.*, 2011 Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. Influenza Other Respir. Viruses 6: 404–416.

Strelkowa, N., and M. Lässig, 2012 Clonal interference in the evolution of influenza. Genetics 192: 671–682.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

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

Wakeley, J., 2008 *Coalescent Theory.* Roberts & Co., Greenwood Village, CO.

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

Communicating editor: J. Hermisson

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.160986/-/DC1

How to Infer Relative Fitness from a Sample of Genomic Sequences

Adel Dayarian and Boris I. Shraiman

DOI: 10.1534/genetics.113.160986

File S1

Supporting Information

Example of a Tree to Describe the Notation

Consider the tree in Figure S1 for a sample of $n = 6$ individuals. The figure shows the label, the time of formation and the weight of each ancestor associated with each internal node. In the main text, we have referred to quantities such as the mean, $F(w_i)$, and the variance, $\bar{\Sigma}^2(w_i)$, in the fitness of the offspring of an ancestor existing some t-generations in the past. As an example, in Figure S1, consider some time t in the past such that $t_3 < t < t_4$, where three lineages exist: 1, l_2 and l_3. The lineage l_3 has weight 2 with descendants 2 and 3 in the sample, while the lineage l_2 hass weight 3 with descendants 4,5,6 in the sample. The mean and the variance in the fitness of the offspring of lineage l_2 is given by $F(3) = \frac{1}{3}(f_4 + f_5 + f_6)$ and $\Sigma^2(3) = \frac{1}{3}((f_4 - F(3))^2 + (f_5 - F(3))^2 + (f_6 - F(3))^2)$, respectively. Similarly, for lineage l_3, the two quantities are given by $F(2) = \frac{1}{2}(f_2 + f_3)$ and $\Sigma^2(2) = \frac{1}{2}((f_2 - F(2))^2 + (f_3 - F(2))^2)$. Finally, for lineage 1 we get $F(1) = f_1$ and $\Sigma^2(1) = 0$. One can repeat a similar procedure for any other time point along the tree.

We now clarify the notation used in the formula for the Fitness Proxy Score. First we need to calculate T_2, the estimate of the average pairwise coalescent time, obtained from the sampled genomes. For the tree in Figure S1, the pairwise coalescent time between lineages 3 and 4 is $\tau_{3,4} = t_4$, or between lineages 1 and 4 is $\tau_{1,4} = t_5$. The average pairwise coalescent time is given by $T_2 = \frac{1}{15}(t_1 + 2 * t_2 + t_3 + 6 * t_4 + 5 * t_5)$. As an example, let us calculate the score of individual 3, ϕ_3. There are $m_3 = 4$ ancestral lineages to individual 3: i) the first line starting from the individual 3 itself, ii) l_3, iii) l_4 and iv) l_5. These four lineages are denoted by a_0 to a_4, respectively. At time t_3, the lineage a_0 with weight 1 merges with

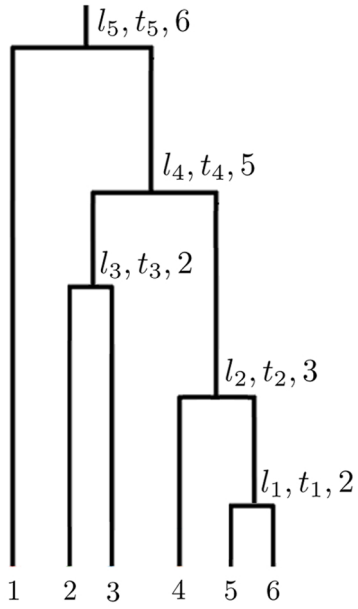

Figure S1: Example of a genealogical tree for a sample of size n=6. The starting 6 lineages carry weight 1 and have fitness values denoted by f_1 to f_6. Lineage l_1 is formed at time t_1 by the merger of two lineages 5 and 6, and carries weight 2. Similarly, lineage l_2 is formed at time t_2 by the merger of two lineages 4 and l_1, and carries weight 3.

another lineage with weight 1 and form a_1 with weight 2. The contribution of this coalescent event to ϕ_3 is equal to $\Theta(t_3/T_2) * (2-1) = \Theta(t_3/T_2) * 1$. Proceeding in a similar fashion, we obtain $\phi_3 = \Theta(t_3/T_2) * 1 + \Theta(t_4/T_2) * 3 + \Theta(t_5/T_2) * 1$. Note that if we ignore the time dependent coefficients $\Theta(t_i/T_2)$, all the individuals get the same score of $n-1=5$.

Evolutionary Simulations

The simulations are done using a custom written Python code, available upon request. The evolution is based on a discrete time Wright-Fisher model with population size N. Each generation t undergoes separate selection and mutation steps. To implement selection, each individual i produces a Poisson-distributed number of gametes in the next generation with parameter $\exp(f_i - \alpha)$. Here $f_i = F_i - \bar{F}$ is the fitness advantage of individual i relative to the mean fitness of the population $\bar{F}$, and $\alpha = \frac{N(t)-N}{N}$ ensures an approximately constant population size around N. Individual genomes are defined as binary strings, g_k with $k = 1, ..., L$ and the number of loci, $L = 10^5$, chosen large enough to exceed the number of segregating

polymorphisms in the simulated population. Consistent with infinite site approximation, new mutations flip the g_a binary value from zero to one.

At each generation, the beneficial and deleterious mutations arise with probability $\epsilon\mu$ and $(1-\epsilon)\mu$ and have a fitness effect of $\pm s$, respectively. We also record the forward genealogies during the simulations. The above process is repeated for a specified number of generations. Statistical measurements on the dynamics of the evolution are taken after an equilibration time to remove transient effects from the initial conditions. In the parameter regimes studied, we found that a "burn in" time of 10^4 generations was generally sufficient.

Given that we perform forward simulations and keep track of the genealogies, the simulations are computationally intensive. Therefore, the maximum population size that we simulated was $N = 64000$. The mutation rate was varied from $\mu = 10^{-4}$ to $\mu = 10^{-2}$ and the selection coefficient from $s = 10^{-3}$ to $8 * 10^{-3}$. Together this spans a $10^{-2} < \mu/s < 10$ range for the all-important μ/s parameter. For the parameter combination where $N = 64000$, $\epsilon = 0.1$ and $\mu = 10^{-4}$ (beneficial mutation rate 10^{-5} and deleterious mutation rate $9 * 10^{-5}$), only a couple of clones are segregating in the population (see below). This parameter combination serves as the boundary between the multisite selection regime and the selective sweep regime. For smaller mutation rates, given that $N = 64000$, the population is monoclonal and enters the regime of selective sweeps.

Below, we present some results on the clonal diversity, as well as the speed of adaptation, for various parameters that we have simulated. In Fig. 2A and B of the main text, we showed two examples of fitness distribution. In Figure S2, we show some more examples. Assume there are c clones in the population, with sizes $n_1, ..., n_c$. Note that $\sum_{i=1}^{c} n_i = N$. To see how many clones with significant size are segregating, we can define the participation fraction: $Y =< \sum_{i=1}^{c} (\frac{n_i}{N})^2 >$. This quantity is equal to the probability that two randomly chosen genomes belong to the same clone. Figure S3A shows the participation fraction ranging from values smaller than 0.001 to values around 0.1 for various sets of parameters. For $N = 64000$ and $\mu = 10^{-4}$, $Y \approx 0.4$, which means that for the smallest value considered for μ, there is a

significant probability that two randomly chosen genomes come from the same clone.

In the regime of our interest, where many mutations simultaneously segregate, it is well known that the competition between mutations slows down the rate of the adaptation (Hill-Robertson or Fisher-Muller effect) (Desai and Fisher, 2007; Rouzine *et al.*, 2008). In Figure S3B, we present the speed of the adaptation, i.e. the rate of change of the mean fitness, normalized by its expected value in the selective sweep regime. In the later regime, the beneficial mutations are rare enough that only a single mutation segregates at a time, and assuming the deleterious mutations are purged, the expected speed of adaptation is $v = 2N\mu\epsilon s^2$. As we see in Figure S3B, for the parameter combination $N = 64000$ and $\mu = 10^{-4}$, the adaptation rate is only around a quarter of its expected value in the selective sweep regime. We see that the normalized speed of adaptation varies from 0.01 to 0.25 in the parameter range that we have considered.

In order to understand the parameter regime in which the interference between mutations becomes significant, consider the following (see (Desai and Fisher, 2007) for further details). Let μ_b denote the beneficial mutation rate, N the population size and s the fitness effect of beneficial mutations. When the population size or mutation rate is small enough, the time it takes for a new mutation to reach xation is less than the time it takes for another new mutation to occur and reach significant size. If a new beneficial mutation reaches the size of order $1/s$ individuals, it will escape the drift with high probability and from that point grows as $\frac{1}{s}\exp(st)$ with time, t. The mutation reaches the size of order $1/s$ individuals in roughly $1/s$ generations and becomes fixated in the order of $\frac{1}{s}\log(Ns)$ generations. Therefore, the total time for a mutation to reach fixation from the initial time of its occurrence is in the order of $\frac{1}{s} + \frac{1}{s}\log(Ns)$ generations. The probability for a mutation to escape drift is roughly s. Since the mutations are generated at rate $N\mu_b$, the time it takes for a mutation destined to escape drift to be generated is $1/(N\mu_b s)$.

If the parameters are such that $N\mu_b s << s/(1 + \log(Ns))$, a new mutation that occurs and sweeps will do so long before the next mutation destined to sweep is occurred. On the

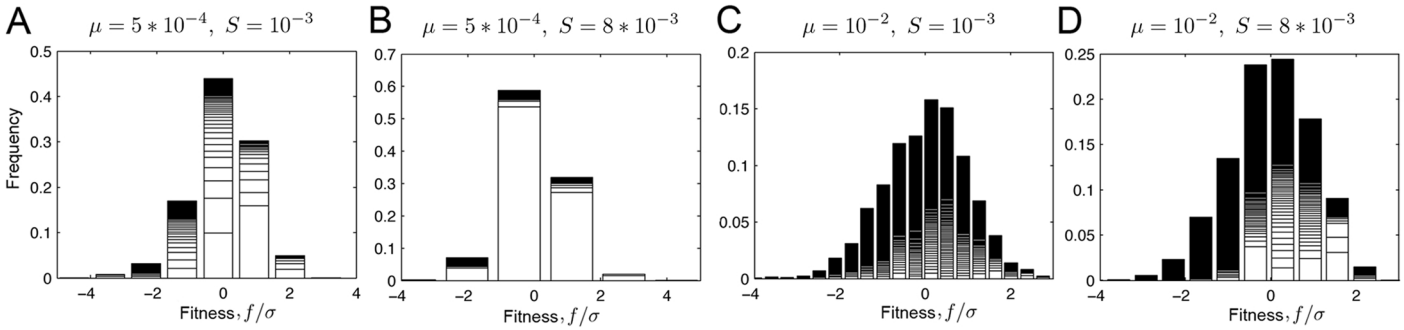

Figure S2: Fitness distribution at one time slice. The mutation rate and selection coefficient for each case is written on the top of the corresponding panel. $N = 64000$ and $\epsilon = 0.1$ for all the panels. Each bin corresponds to a fitness class and each class is composed of several clones. The height of each box within each bin represents the size of a clone. Larger clones are stacked on the bottom. The dark band on top of each bin correspond to small clones.

other hand, for higher mutation rates or population sizes where the above inequality is not satisfied, new beneficial mutations arise and reach significant size before earlier ones can sweep, causing them to interfere with one another. For the case of purifying selection where only deleterious mutations are present ($\epsilon = 0$), it can be shown (Walczak *et al.*, 2012) that the required condition is $N \exp(-\mu_d/s) << \frac{1}{s}\log(\mu_d/s)$, where μ_d is the deleterious mutation rate and s is the deleterious effect of mutations. For example, later we will show results using the parameter combination $N = 32000$, $\mu = 5 * 10^{-3}$, $s = 10^{-3}$ and $\epsilon = 0$. For this combination, we have $N \exp(-\mu_d/s) \simeq 216$ and $\frac{1}{s}\log(\mu_d/s) \simeq 1609$. For smaller μ_d or higher s such that the above condition is not satisfied, the deleterious mutations are purged out of the population fast enough such that the effect of selection can be captured by a simple effective population-size approximation.

Tree Reconstruction

In the first step, we use the neighbor-joining algorithm (Durbin, 1998; Felsenstein, 2004) to reconstruct the tree topology. The input distance matrix for this algorithm is simply given by the pairwise difference of sequences (Hamming distance) including both neutral and non-neutral mutations. The time to the common ancestor of two individuals is proportional to the number of neutral genetic differences between them. For a real data set, one may use a

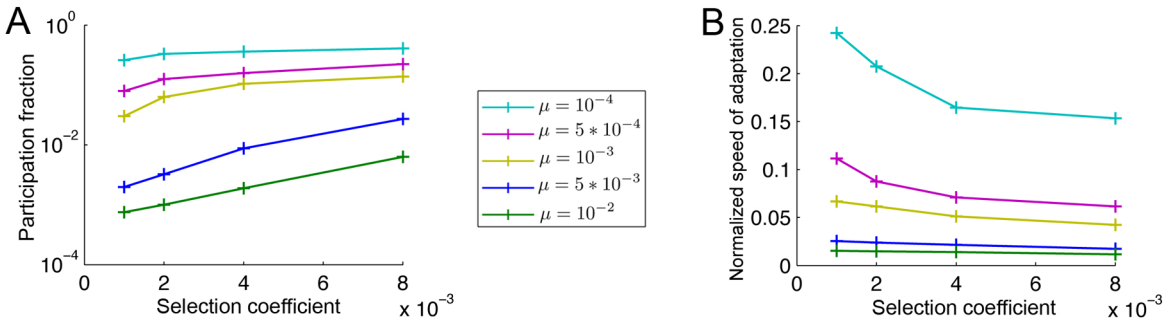

Figure S3: Clonal structure and adaptation rate. (A) The participation fraction, Y, for different parameter values. Y is the probability that two randomly chosen genomes belong to the same clone. For all the curves, $N = 64000$ and $\epsilon = 0.1$. When $\mu = 10^{-4}$ (beneficial mutation rate 10^{-5} and deleterious mutation rate $9*10^{-5}$), the values of Y become significant (> 0.1). This implies that the dynamics is at the boundary between the multisite selection regime and the selective sweep regime. (B) Speed of adaptation, normalized by its expectation value in the limit of selective sweep $2N\mu\epsilon s^2$.

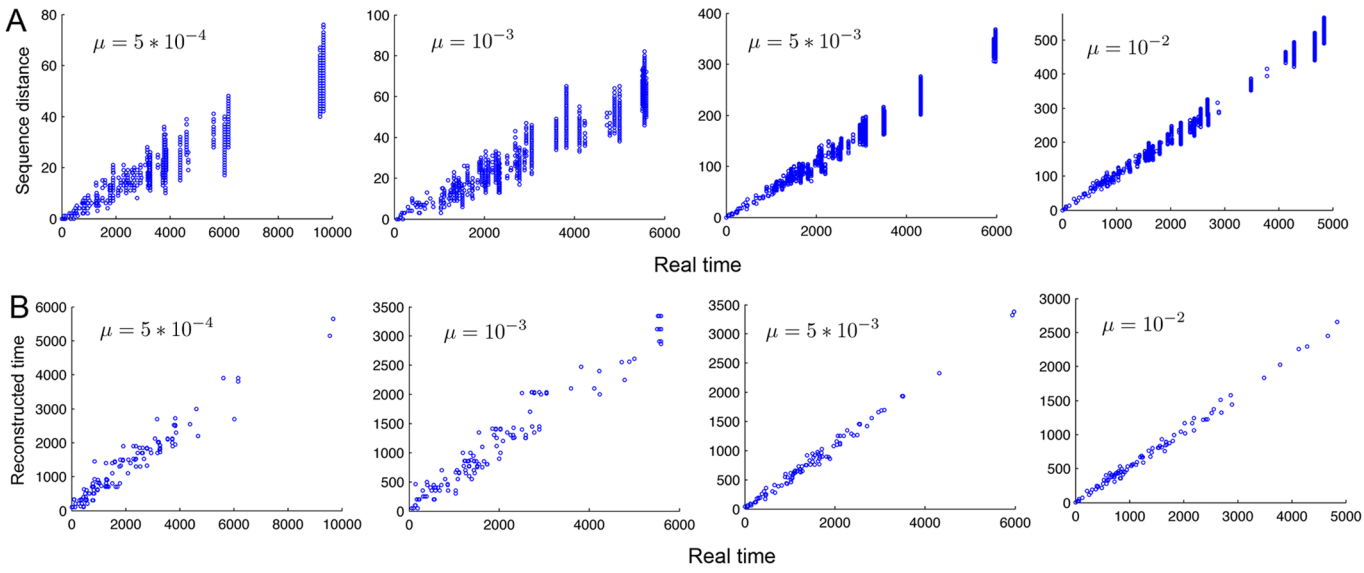

Figure S4: Examples of tree reconstruction for sample size of $n = 100$. $N = 64000$, $\epsilon = 0.1$ and $s = 2 * 10^{-3}$ in all cases. (A) Scatter plot of the Hamming distance between sequences versus the real divergence time for all the pairs in the sample. The non-neutral mutation rate for each case is shown in the associated plot. The neutral mutation rate was set to 10 times the value of the non-neutral mutation rate. (B) Scatter plot of the reconstructed divergence time between sequences versus the real divergence time for all the pairs in the sample. Each plot is associated to the same tree as in panel (A) for the same mutation rate.

more realistic substitution model to infer the divergence time between pairs of genomes. We have considered the values of the non-neutral mutation rate over a few orders of magnitude. The neutral mutation rate was always set to 10 times the value of the non-neutral mutation rate. We use the neighbor-joining algorithm only to infer the topology of the tree. We do not use the length of the edges that are calculated in this algorithm. The reason is that we want all the leaves of the tree to be located at the current time and have the same distance to the root.

In the next step we find the root of the tree based on the parsimony method. Each point on the tree divides the sample into two groups. The root should be located at a point where the similarity between the two groups is minimal. We count the number of mutations which exist in both groups and assign the root to a point where this number is minimal.

In the last step, we assign the height (time interval to the present time) of each node in the tree. The lengths are calculated as in the UPGMA algorithm (Felsenstein, 2004). In this algorithm, the total branch length from a tip down to any node is half of the average of the distance between all the pairs of genomes whose most recent common ancestor is that node. We consider a node only after all the nodes below it have their heights assigned. We start from the bottom, namely, the nodes which are connected to two leaves. The height of these nodes are calculated similar to the UPGMA algorithm: the height is equal to the half of the mutational distance between the pair of the genomes below that node. For other internal nodes, we also calculate the putative height as half of the distance between all the pairs whose most recent common ancestor is that node. The height of the node is the maximum between this putative distance and the height of all the internal nodes below the considered node.

We evaluated the performance of the above tree reconstruction algorithm in all different parameter ranges by comparing the reconstructed tree with the actual genealogy. In all the cases, the performance was satisfactory. In Figure S4, we show examples of the performance of the above algorithm for four different mutation rates. For each rate, a sample of size

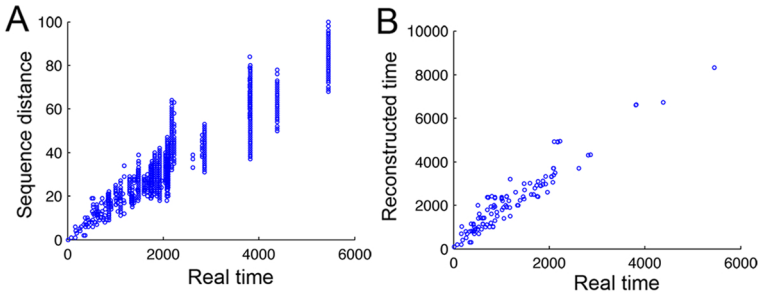

Figure S5: Examples of tree reconstruction for sample size of $n = 100$. $N = 64000$, $\epsilon = 0.1$, $s = 2 * 10^{-3}$ and $\mu = 10^{-2}$ in both plots. The neutral mutation rate was set equal to its value for the non-neutral mutation rate, μ. (A) Scatter plot of the Hamming distance between sequences versus the real divergence time for all the pairs in the sample. (B) Scatter plot of the reconstructed divergence time between sequences versus the real divergence time for all the pairs in the sample.

$n = 100$ is selected. In Figure S4A, we show the sequence distance for all the $(n-1)n/2$ pairs in the sample versus the real divergence time. These distances are the input of the above algorithm. In Figure S4B, we show the reconstructed divergence time (inferred from the reconstructed tree) for all the pairs. The validity of the above algorithm is reflected in the fact that the relation between these two times is close to being linear. The slope of the line is irrelevant, since, it only reflects an scaling factor, i.e. the estimation of the mutation rate. As we see in Figure S4, for higher mutation rates (e.g. $\mu = 10^{-2}$) where a typical pair of genomes are polymorphic at hundreds of loci, neutral mutation rate need not be 10 time that of μ for the tree reconstruction to be accurate. In these cases, there is enough diversity that even setting the neutral mutation rate equal to μ would be sufficient (see Figure S5).

Weight Distribution

Consider a sample of size n and the corresponding phylogenetic tree. Assume looking at the tree at the stage where there are a lineages left. The ancestor i will carry a weight w_i where $i = 1, ..., a$ and $\sum_{i=1}^{a} w_i = n$. The values that w_i can take is anything between 1 and $n - a + 1$. For example, when there are only 2 ancestral lineages, w_i can be between 1 and $n - 1$. The statistics of the phylogenetic trees for neutral evolution are given by the Kingman's coalescent (Kingman, 1982a,b). In particular, the probability distribution of w_i is given by (Derrida

and PELITI, 1991):

$$P_{neu}(w_i|a,n) = \binom{n-w_i-1}{a-2} \Big/ \binom{n-1}{a-1} \tag{1}$$

For example, when there is only $a = 2$ ancestors left in the tree, we get $P_{neu}(w|2,n) = \frac{1}{n-1}$, which is independent of w. The above formula can be derived solely based on the fact that, as one goes up in the tree, at each stage, any lineage is equally likely to coalesce with any other lineage regardless of the weight they are carrying or any other previous events in the tree.

Distortion in Shape of Genealogical Trees

Here, we consider some quantities which reflect the differences between the shape of trees from non-neutral and neutral evolution. While inspecting trees in Fig. 2D and E, we notice that in the presence of selection it is more common for a leaf (sampled genome) to be connected to a long edge. In other words, it takes a relatively long time for some leaves to merge to other lineages in the tree. Moreover, such leaves are more likely to belong to lower fitness classes, represented by blue and grey colors. In addition, number of lineages left in a tree as a function of time seems to be different. Here, we explore such points in more details.

At each instant of the time in a tree, one can consider what fraction of the remaining lineages are singletons. Singletons are defined as lineages with weight *w=1*. In Figure S6A shows the average value of this fraction as a function of time. These curves are obtained by averaging over random samples and over population replicas. The time for each tree is linearly rescaled so that the current time is at 0 and the root is at 1. At time 0, all the lineages in a tree are singletons and the fraction is, therefore, one. At time 1, all the lineages have merged together and therefore no singleton lineage is left. As we see, the curve for the neutral case falls below the rest of the curves. In the neutral case, the statistic of length of singleton edges was studied in (FU and LI, 1993) and is used in the Fu and Li's test for

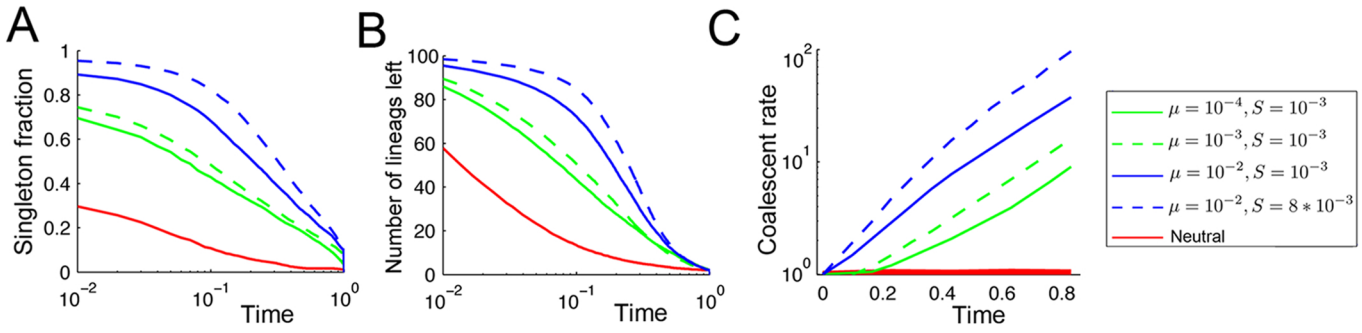

Figure S6: Statistics on the time dependence of number of lineages in phylogenetic trees. For all the panels, the sample size is $n = 100$, $N = 64000$ and $\epsilon = 0.1$. (A) Average fraction of singleton lineages left in a tree as a function of time. The time for each tree has been linearly rescaled so that the root is at $t = 1$ and the current time is 0. (B) Average number of lineages left in a tree as a function of time. The time has been linearly rescaled as in part (A). (C) Coalescent rate between two random lineages as a function of time. The rate is normalized by its value at time $t = 0$. The 'effective population size', N_e, would be defined to be inversely proportional to coalescent rate.

detecting departures from the Kingman's coalescent.

One can also study the fitness of the singleton lineage that is the latest to join the rest of the tree. Figure S7 shows the fitness distribution using the same simulation parameters as in Fig. 2D. As we see, these lineages tend to belong to the unfit classes. This is a general pattern observed for all of the simulation parameters.

We have also considered the average number of lineages left in a tree as a function of time, $< a >_t$. The result is presented in Figure S6B. In the presence of selection, the number of lineages drops slower at early times compared to the neutral case. Under neutral evolution, since the coalescent events happen at rate $\binom{a}{2}/N$, when there are a lineages left, one has: $\frac{d<a>_t}{dt} \propto - < \binom{a}{2}/N >$. Therefore, the ratio $-\frac{d<a>_t}{dt} / < \binom{a}{2}/N >$ which is the coalescent rate between two random lineages remains constant. Figure S6C shows the coalescent rate normalized by its value at time $t = 0$. For the neutral case, this rate remain constant, as expected. However, in the presence of selection, the rate increases for further time back in the tree. The reason for this is that, for times further back in the tree, the ancestral lineages are more likely to have belonged to the leading edge of the fitness distribution at the time they existed (see Fig. 4A of the main text). Therefore, they coalesce at a faster rate compared

to the bottom of the tree where lineages are spread over the fitness distribution (O'Fallon *et al.*, 2010).

Increase in the coalescent rate is sometimes interpreted as a reduction in the effective population size (denoted by N_e). However, not all aspect of the coalescent process under selection, such as the weight distribution or fraction of singleton lineages, can be accounted for simply by introducing an effective population size. This fact also manifests itself in the distribution of polymorphisms in a sample of genomes. Under neutrality (in the limit of infinite-site model), the probability that a derived allele appears in w individuals out of n sampled genomes is proportional to $1/w$. This behavior is a consequence of both the weight distribution and the length of coalescent intervals. To see this, note that in order to appear in w genomes in the sample, a mutation must have occurred on an ancestor with weight w. Assume this ancestor existed when there was a lineages in the tree. The probability that 1 of the a ancestors carried weight w is $a * P_{neu}(w|a,n)$. The average time a tree spends having a lineages is proportional to $1/\binom{a}{2}$. Summing over all possible a's gives: $\sum_{a=2}^{n} aP_{neu}(w|a,n)\frac{1}{\binom{a}{2}} = \frac{2}{w}$, which is the usual one over frequency dependence. In Figure S8 we show the frequency distribution of neutral polymorphisms in the presence of selection. The distribution first drops more like $1/w^2$ for small frequencies and then bends upward for higher frequencies where $w > n/2$ (Neher and Shraiman, 2011; Neher, 2013).

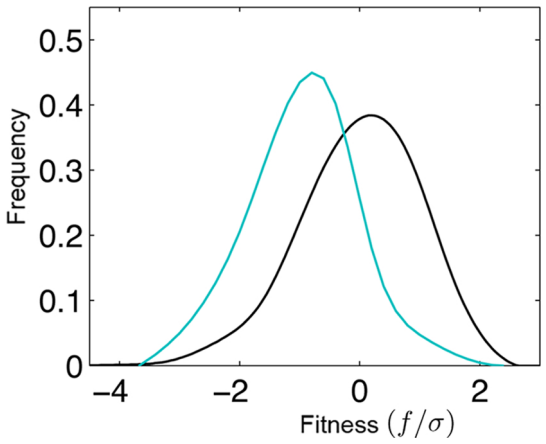

Figure S7: Fitness distribution of the singleton lineage which connects to the tree the latest is shown in cyan color. The black curve presents the fitness distribution of the whole population which is the same as fitness distribution of the sampled genomes. The distributions are obtained by averaging over random samples and over population replicas. $N = 64000$, $\mu = 10^{-3}$, $s = 2 * 10^{-3}$ and $\epsilon = 0.1$.

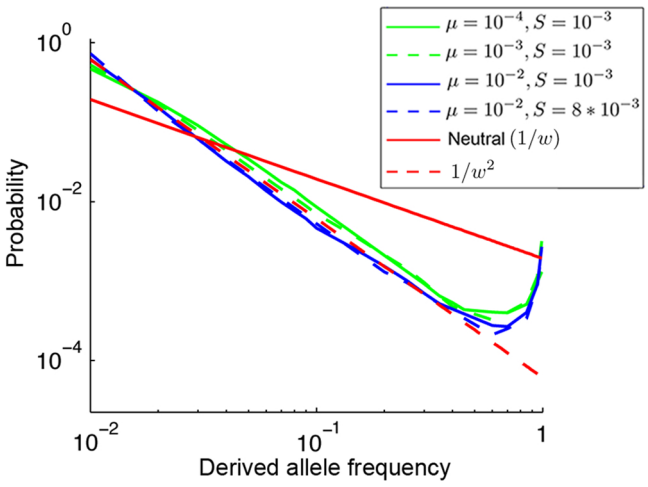

Figure S8: Distribution of neutral polymorphism in a sample of size $n = 100$. The dashed red line shows the probability distribution which dependents on $1/w^2$, as opposed to $1/w$ (neutral case). $N = 64000$, $\epsilon = 0.1$.

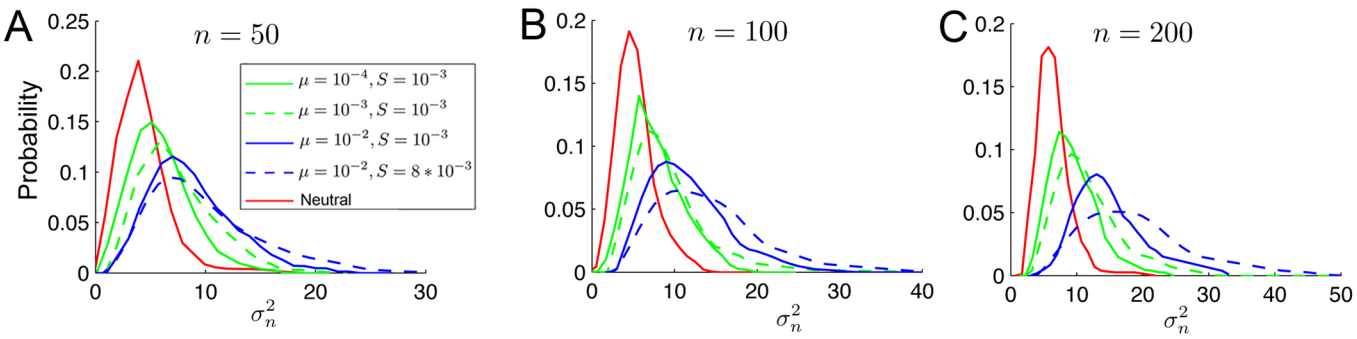

Figure S9: Distribution of σ_n^2, a measure of asymmetry based on the topology, for three samples sizes of $n = 50$, $n = 100$ and $n = 150$. In all cases, $N = 64000$ and $\epsilon = 0.1$.

Test of neutrality based on the shape of trees

We discuss some measures for distinguishing between neutral and non-neutral trees. Let us use the term topology to refer solely to the branching pattern of a tree. On the other hand, the term shape will refer to the information about both the branching pattern and the branch lengths. In Kirkpatrick and Slatkin (1993), authors reviewed six measures of tree asymmetry based solely on the tree topology. They studied the power of these measures to be used as a test for deviation of trees from neutral predictions. A similar analysis was carried out in Maia *et al.* (2004). One of the measures denoted by σ_n^2 was identified as the relatively more powerful in both studies. Below, we show the result of applying this measure to the trees obtained in simulations.

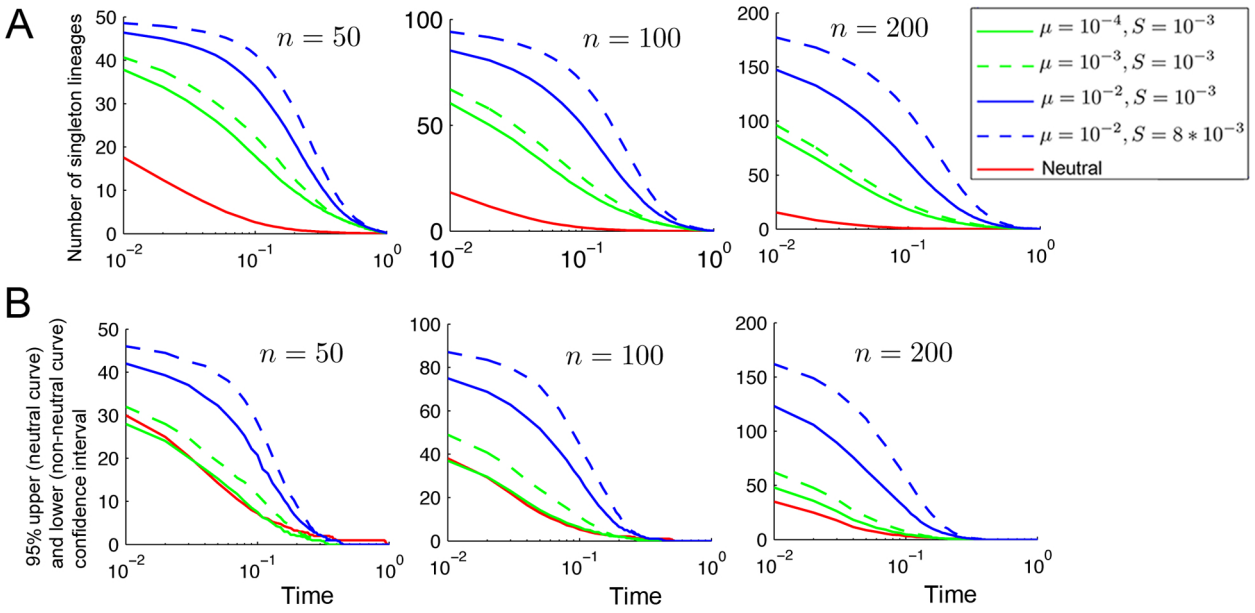

Figure S10: (A) The average number of singleton lineages left in a tree as a function of time, for three samples sizes of $n = 50$, $n = 100$ and $n = 150$. The time has been linearly rescaled so that the root is at $t = 1$ and the current time is 0. (B) The upper (for the neutral curve) and lower (for the non-neutral curves) %95 confidence intervals for the curves in part A. In all cases, $N = 64000$, $\epsilon = 0.1$.

To each leaf i in a tree, a number N_i is assigned. This is the number of internal nodes between leaf i and the root. The variance of this number in a tree is given by $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(N_i - \bar{N})^2$. In a completely symmetric tree $\sigma_n^2 = 0$. Figure S9A presents the distribution of σ_n^2 for both neutral and non-neutral trees for three sample sizes. As expected, the results indicates that the coalescent trees in the presence of selection are, on the average, more asymmetric compared to neutral trees. In addition, as the sample size increases, the distribution of σ_n^2 differs more between the neutral and non-neutral cases. However, even for sample size $n = 200$, there is a significant overlap between the neutral and non-neutral distributions. Therefore, this measure is not a useful test to detect a tree signicantly distinct from the neutral expectation. Similar conclusion was reached in Maia *et al.* (2004) where authors have analyzed three more measures than the one presented here.

The above measure does not take into account the information about the branch-length in a tree. We have looked at two quantities which use this additional information. Figure S10A

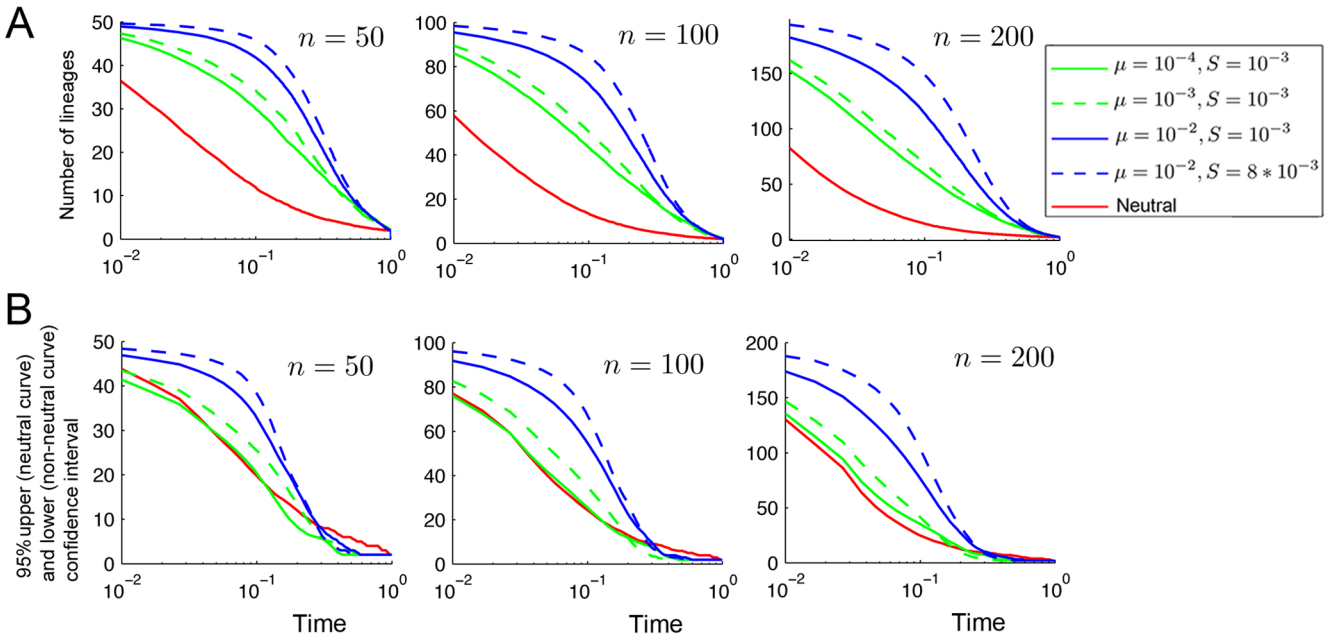

Figure S11: (A) The average number of lineages left in a tree as a function of time, for three samples sizes of $n = 50$, $n = 100$ and $n = 150$. The time has been linearly rescaled so that the root is at $t = 1$ and the current time is 0. (B) The upper (for the neutral curve) and lower (for the non-neutral curves) %95 confidence intervals for the curves in part (A). In all cases, $N = 64000$, $\epsilon = 0.1$

shows the expected number of singleton lineages left in a tree as a function of time for three sample sizes. These curves are obtained by averaging over random samples and over population replicas. The curve for the neutral case falls clearly below the rest of the curves. To see if the separation between the neutral and non-neutral curves is large enough that one can differentiate whether or not a single tree is neutral, we also show the confidence intervals in Figure S10B. The upper 95% confidence interval for the neutral case falls below the lower %95 confidence for almost all the cases with selection. The separation becomes larger as the sample size increases. Even for the parameter combination where the dynamics falls at the boundary between the multisite selection and the selective sweep regime ($N = 64000$, $\mu = 10^{-4}$, $s = 10^{-3}$ and $\epsilon = 0.1$), the lower confidence interval is very close to the upper confidence interval for the neutral curve. Figure S11A shows the average number of lineages left in a tree as a function of time. The confidence intervals are also shown in Figure S11B. Again, for $n = 100$ or $n = 200$, the upper confidence interval curve for the neutral case falls

below the lower confidence interval curves for all the cases with selection.

Correlation between Weight and Fitness of Ancestors

In Fig. 4A of the main text, we showed the distribution of the fitness of the ancestors for certain time intervals in the past for a set of parameters. Let us denote this distribution by $D_t(f)$. In the limit of large times, this distribution is equal to the fitness distribution for the common ancestor of the whole population, $D_\infty(f)$. In Fig. 4B, we also showed the scatter plot between the weight of ancestors and their fitness advantage for $t = 100$ generations in the past. The scatter plot represents the joint distribution of weight and fitness of ancestors, $D_t(f, w)$.

One can consider the expected fitness of an ancestor given its weight, $\bar{f}_{anc}(w,t) = \sum_f f * D_t(f|w)$. Figure S12A shows $\bar{f}_{anc}(w,t)/\sigma$ as a function of w/N for $t = 100$ and $t = 500$ in log-log scale. The dependence seems to be linear, namely, $\bar{f}_{anc}(w,t) \propto w^{m(t)}$, where $m(t)$ is the slope of the lines in Figure S12A. This slope depends on the time, and of course, other parameters such as N, μ, etc. Figure S12B shows $m(t)$ as a function of time for different sets of parameters. For each set of parameters, the time axis has been rescaled with the fitness variance for the corresponding parameter set, $\sigma(N, \mu, \epsilon, s)$. As we see, the slope $m(t)$ drops as a function of time. In other words, the correlation between the weight and the fitness of ancestors reduces as one goes further back in time.

In the main text, we also presented some results on the relation between the weight of an ancestor in a tree, w_i, and the fitness of the w_i's genomes in the sample which are derived from that ancestor. In particular, we focused on the mean, $F(w_i) = \frac{1}{w_i}\sum_{j=1}^{w_i} f_j$, and the variance, $\Sigma^2(w_i) = \frac{1}{w_i}\sum_{j=1}^{w_i}(f_j - F(w_i))^2$ (see below for an example of a tree explaining the notation). The average of these quantities over random samples of genomes and over population replicas are denoted by $\bar{F}(w_i) =< \frac{1}{w_i}\sum_{j=1}^{w_i} f_j >$ and $\bar{\Sigma}^2(w_i) =< \frac{1}{w_i}\sum_{j=1}^{w_i}(f_j - F(w_i))^2 >$.

In the main text, we only presented these quantities for two parameter sets. In Figure S13A and B, we show $\bar{F}(w)/\sigma$ and $\bar{\Sigma}^2(w)/\sigma$ for more parameter sets. The sample size is

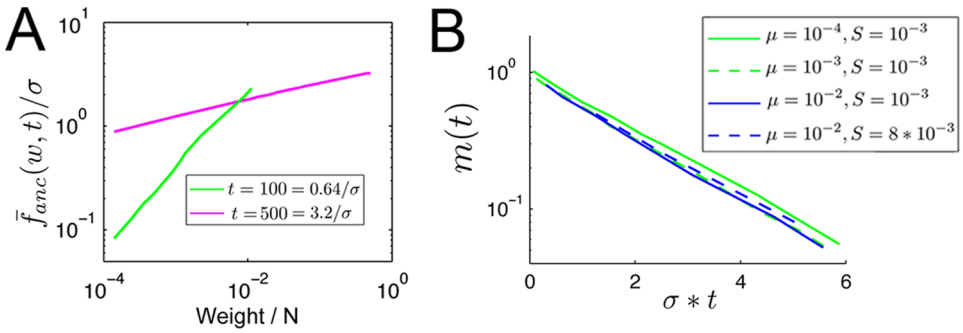

Figure S12: Correlation between the weight and fitness of ancestors. (A) Average fitness of an ancestor conditional on its weight for two time intervals. Note the log-log scale. $N = 64000$, $\epsilon = 0.1$, $\mu = 10^{-3}$ and $s = 8 * 10^{-3}$. (B) Fitting a line to the curve in part (A) gives a time dependent slope $m(t) = \log(\bar{f}_{anc}(w,t))/\log(w)$. The slope $m(t)$ is plotted as a function of time for a few different parameter values. Note the log scale on the y-axis. The time for each parameter set has been rescaled by the corresponding σ. Sample size $n = 100$, $N = 64000$ and $\epsilon = 0.1$.

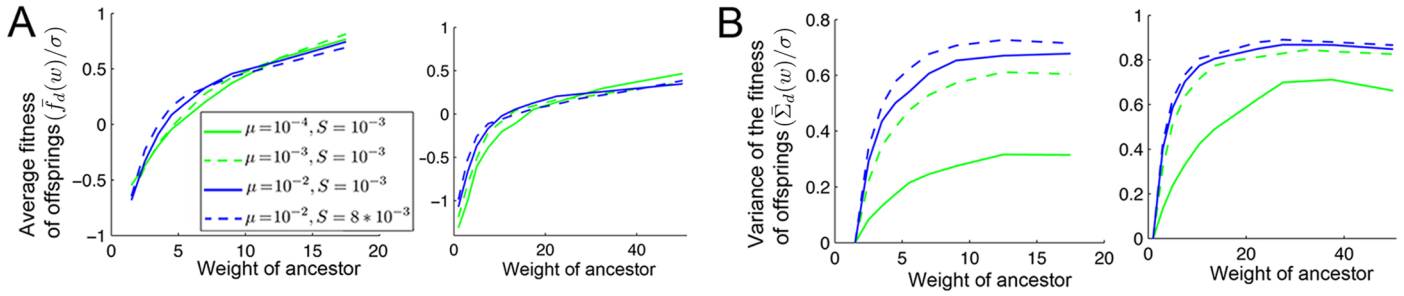

Figure S13: Correlation between the weight and fitness of offspring. (A) Average fitness of genomes as a function of the ancestral weight for two different time slices in the past. (B) Variance in the fitness of genomes as a function of the ancestral weight for two different time slices in the past. $N = 64000$, $\epsilon = 0.1$ in all cases.

$n = 100$ and the results are shown for two different time points. One of the time points is chosen to be the first time that the tree carries a lineage with weight greater than 15% of the sample size. The other time point corresponds to the first time the weight of a single lineage becomes greater than 40% of the sample size.

Fitness Proxy Score and its Performance

Consider a sample of n genomes and the corresponding reconstructed phylogenetic tree. Although there is always a positive correlation between the weight of an ancestor and its fitness and the fitness of its derived genomes, both of these correlations drop as one goes further back in time. When most of the lineages have condensed into high-weight ancestors, the average fitness of the offspring of such ancestors is close to zero and there is little correlation between the weight and the fitness of the offspring (see right plot in Fig. 4C of the main text). The variance in the fitness of the offspring also becomes close to the population fitness variance σ. In other terms, all of the derived genomes of such high-weight ancestors are, more or less, evenly distributed across the fitness distribution.

This is consistent with our observations in Fig. 4D of the main text. As the coalescent time for a pair of genomes increase, the difference in the fitness of the two genomes increases as well. This means, as τ_{ij} becomes larger compared to T_2 (the region covered in yellow and red colors in Fig. 4D), there is less information about the fitness of the pair of genomes involved. For example, one can have high fitness and the other one low fitness, or both can have average fitness. In other words, when the coalescent time for a pair of genomes becomes larger compared to the population average T_2, there is more uncertainty on the fitness of that pair of genomes.

Because of the above argument, we do not want the scoring scheme to be affected by the coalescent events far back in the tree. In addition, as we saw in Fig. 4D of the main text, when the fitness of two genomes is higher, the coalescent time between them is shorter compared to the mean pairwise coalescent time for the whole population T_2. The correlation between

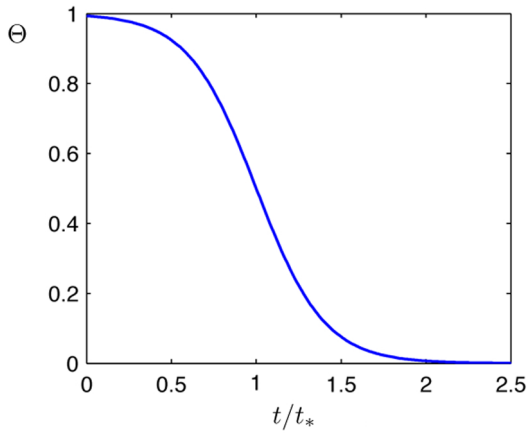

Figure S14: The Fermi-Dirac function $\Theta(t) = \left(1 + \exp(5 \times (t/t_* - 1))\right)^{-1}$.

weights and fitness is also stronger for earlier times. Therefore, the earlier a coalescent event, the more it should affect the scores. It is important to have a sense of 'early times' or 'late times' in a tree. We use the empirical value of the mean pairwise coalescent time (i.e. estimate of T_2 from the sample) for this purpose. In the algorithm, the time values appear only in the form of ratios. So having a correct estimate of the mutation rate is irrelevant.

To incorporate the above ideas, we introduced a threshold time $t_* = x_* \times T_2$ and have the coalescent events which happen at a time further back compared to t_* contribute progressively less to the score. On the other hand, the coalescent events earlier than this stage will be progressively more important in the scoring scheme. In order to do this, we introduced the function $\Theta(t)$ with a Fermi-Dirac form, shown in Figure S14. In the results shown in this paper on the performance of the algorithm, we set $t_* = 0.5 * T_2$, where T_2 is the average pairwise coalescent time. We checked the performance for various values of t_*. We found that, in general, the results are very robust within a range of $0.4 * T_2 < t_* < T_2$. Outside this range the performance slightly decreases. For the sake of example, in Figure S15, we show the probability for the fitness of a genome within the top %10 ranked to belong to the top 50% fitness values as a function of the threshold parameter, x_*, for two different mutation rates.

We have also evaluated the performance of the algorithm for different sample sizes. In

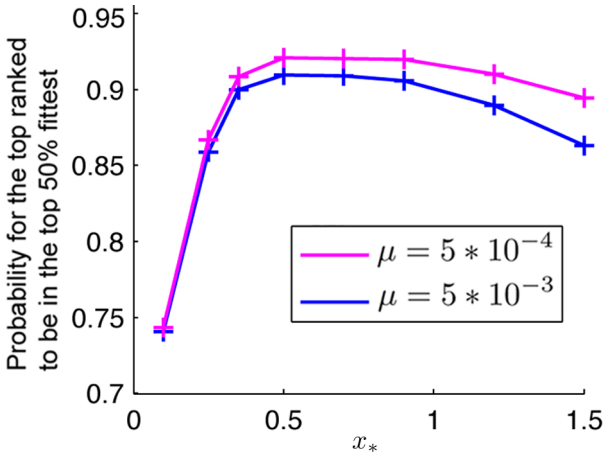

Figure S15: Performance as a function of the threshold x_*. $N = 64000$, $\epsilon = 0.1$ and $s = 2 * 10^{-3}$.

Figure S16, we present the results for a set of parameters. We see that for samples smaller than $n = 100$. the performance decreases, whereas for higher samples sizes, the performance is similar to the results shown above for $n = 200$. For example, for a sample of size $n = 30$, and for parameters $\mu = 5 * 10^{-3}$ and $s = 2 * 10^{-3}$, the probability for the fitness of the top ranked genome to belong to the top 50% values turns out to be around 0.84, compared to 0.9 for sample size of $n = 200$.

Another point is that, as sample size becomes smaller, the right tail of the fitness distribution (see Fig. 2A and B) becomes under sampled. It has been shown that in similar models as the one we have considered here, the bulk of the fitness distribution can be approximated by a Gaussian profile (Desai and Fisher, 2007). For a Gaussian distribution, the probability of sampling a point with value of at least one (two) σ above the mean is around 0.15 (0.3). By inspecting the fitness profiles in Fig. 2A and B of the main text, as well as profiles shown in Figure S2 of SI, we see that the frequency of clones with fitness more than one σ above the population average is around 0.1. This frequency for clones with fitness more than 2σ above the population average is less than $p = 0.05$. In Figure S16C, we see the ratio of the maximum fitness value in a sample of size n to the maximum fitness value that exist in the population. As expected, the larger the sample size, this ratio gets closer to one.

We have tested the performance of the algorithm for the case of high mutation rates with

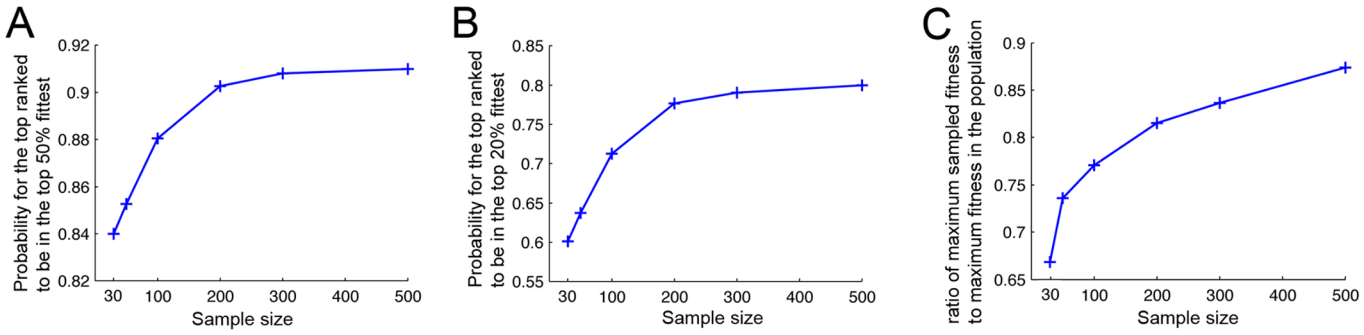

Figure S16: Performance as a function of the sample size. $N = 64000$, $\epsilon = 0.1$, $\mu = 5 * 10^{-3}$ and $s = 2 * 10^{-3}$.

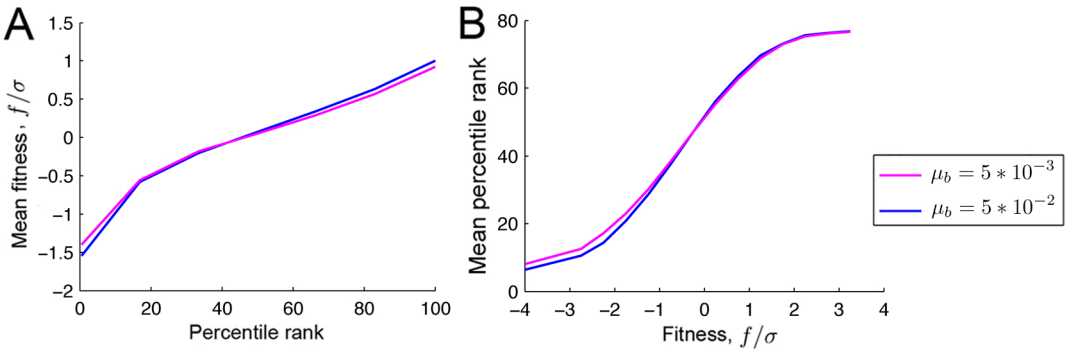

Figure S17: Performance of the fitness ranking algorithm for high beneficial mutation rates. Sample size $n = 200$, $N = 64000$, $\epsilon = 1$ and $s = 2 * 10^{-3}$ in all plots. (A) Mean fitness as a function of the rank for $\mu = 5 * 10^{-3}$ and $\mu = 5 * 10^{-2}$. (B) Mean rank as a function of the fitness for $\mu = 5 * 10^{-3}$ and $\mu = 5 * 10^{-2}$.

only beneficial mutations present (i.e. $\epsilon = 1$). The results shown in Figure S17 indicate that the algorithm performs well in this regime. We have also studied the performance of the algorithm in the presence of purifying selection. i.e. when $\epsilon = 0$. The results presented in Figure S18 show that the algorithm performs well in this case, similar to the case where both beneficial and deleterious mutations are present.

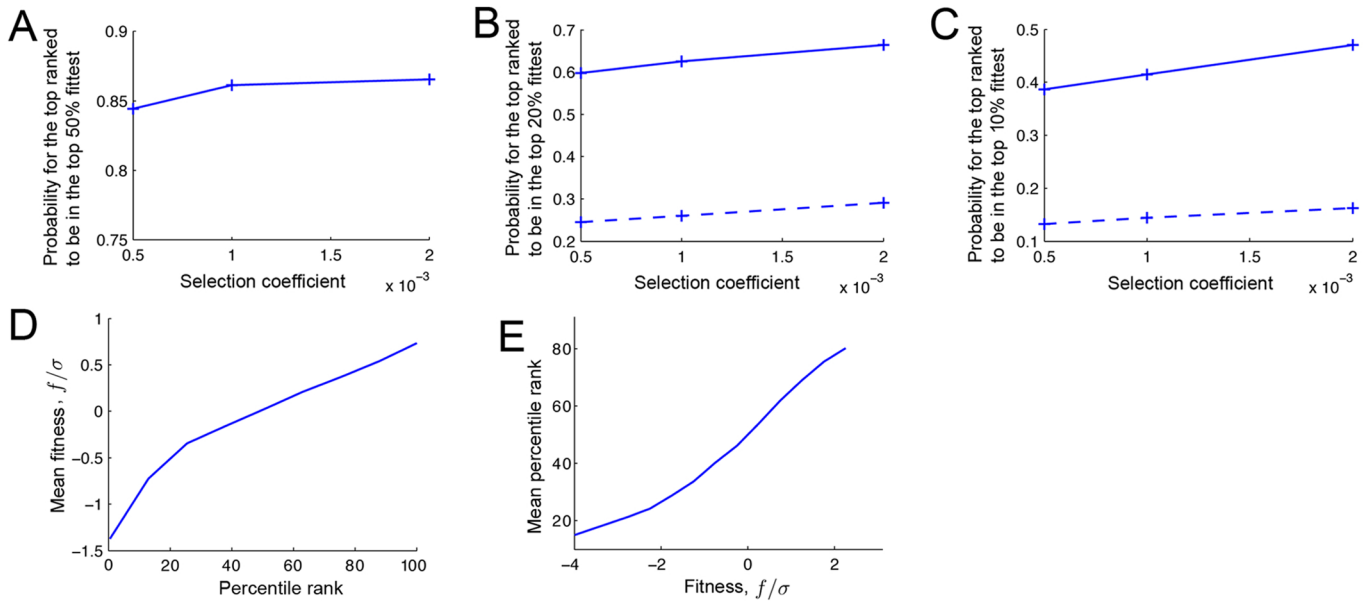

Figure S18: Performance of the fitness ranking algorithm in the case of purifying selection. Sample size $n = 200$, $N = 32000$ and $\mu = 5 * 10^{-3}$ in all plots. (A) Probability for the fitness of a genome within the top 10% ranked to belong to the top 50% fitness values. (B) Probability for the fitness of a genome within the top 10% ranked to belong to the top 20% fitness values. The dashed line shows this probability for a randomly chosen genome. (C) Probability for the fitness of a genome within the top 10% ranked to belong to the top 10% fitness values. (D) Mean fitness as a function of the rank. Selection coefficient $s = 10^{-3}$. (E) Mean rank as a function of the fitness. Selection coefficient $s = 10^{-3}$.

References

Derrida, B., and L. Peliti, 1991 Evolution in a flat fitness landscape. Bulletin of mathematical biology **53**: 355–382.

Desai, M., and D. Fisher, 2007 Beneficial mutation–selection balance and the effect of linkage on positive selection. Genetics **176**: 1759.

Durbin, R., 1998 *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr.

Felsenstein, J., 2004 Inferring phylogenies. Sunderland, Massachusetts: Sinauer Associates .

Fu, Y., and W. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133**: 693–709.

Kingman, J., 1982a The coalescent. Stochastic processes and their applications **13**: 235–248.

Kingman, J., 1982b On the genealogy of large populations. Journal of Applied Probability : 27–43.

Kirkpatrick, M., and M. Slatkin, 1993 Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution : 1171–1181.

Maia, L., A. Colato, and J. Fontanari, 2004 Effect of selection on the topology of genealogical trees. Journal of theoretical biology **226**: 315–320.

Neher, R., and B. Shraiman, 2011 Genetic draft and quasi-neutrality in large facultatively sexual populations. Genetics **188**: 975–996.

Neher, R. A., 2013 Genetic draft, selective interference, and population genetics of rapid adaptation. arXiv preprint arXiv:1302.1148 .

O'Fallon, B., J. Seger, and F. Adler, 2010 A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. Molecular biology and evolution **27**: 1162–1172.

Rouzine, I., É. Brunet, and C. Wilke, 2008 The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. Theoretical population biology **73**: 24–46.

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