

The Dynamics of Genetic Draft in Rapidly Adapting Populations

Katya Kosheleva and Michael M. Desai¹

Department of Organismic and Evolutionary Biology, Department of Physics, and Faculty of Arts and Sciences Center for Systems Biology, Harvard University, Cambridge MA 02138

ABSTRACT The accumulation of beneficial mutations on competing genetic backgrounds in rapidly adapting populations has a striking impact on evolutionary dynamics. This effect, known as clonal interference, causes erratic fluctuations in the frequencies of observed mutations, randomizes the fixation times of successful mutations, and leaves distinct signatures on patterns of genetic variation. Here, we show how this form of “genetic draft” affects the forward-time dynamics of site frequencies in rapidly adapting asexual populations. We calculate the probability that mutations at individual sites shift in frequency over a characteristic timescale, extending Gillespie’s original model of draft to the case where many strongly selected beneficial mutations segregate simultaneously. We then derive the sojourn time of mutant alleles, the expected fixation time of successful mutants, and the site frequency spectrum of beneficial and neutral mutations. Finally, we show how this form of draft affects inferences in the McDonald–Kreitman test and how it relates to recent observations that some aspects of genetic diversity are described by the Bolthausen–Sznitman coalescent in the limit of very rapid adaptation.

THE effects of linkage between beneficial mutations in altering evolutionary dynamics and the structures of genealogies in adapting populations has been recognized for nearly a century, particularly in the context of the evolutionary advantage of sex (Muller 1932). In both asexually reproducing organisms and in regions of low recombination in sexual organisms, the chance congregation of beneficial mutations on competing genetic backgrounds skews evolutionary dynamics. Because of this “clonal interference” effect, the success of a mutation depends not only on its fitness effect, but also on the quality of the genetic background in which it occurs and the fortune of the mutant’s progeny in amassing more beneficial mutations (Smith and Haigh 1974; Gerrish and Lenski 1998; Gillespie 2000, 2001; Kim and Orr 2005).

Recent work in experimental evolution has confirmed that clonal interference is widespread in large adapting laboratory microbial and viral populations (de Visser *et al.* 1999; Miralles *et al.* 1999; de Visser and Rozen 2006; Kao

and Sherlock 2008; Lang *et al.* 2011, 2013). Several recent studies also suggest that classic “hard” selective sweeps may be rare in *Drosophila* (Sella *et al.* 2009; Karasov *et al.* 2010) and humans (Pritchard *et al.* 2010; Hernandez *et al.* 2011), implying that models that better account for linkage between sites need to be explored. As a result, in recent years there has been an influx of theoretical work describing the effects of clonal interference on the evolution of large populations (see Park *et al.* 2010 for a recent review).

This work has provided a good understanding of evolutionary dynamics in the regime of *rare* interference, where the number of strongly beneficial mutations segregating in a population is rarely more than two (Gerrish and Lenski 1998; Gillespie 2000, 2001; Kim and Stephan 2003; Park and Krug 2007). However, in large populations many beneficial mutations can segregate simultaneously and the population can maintain substantial variation in fitness. This decreases the importance of each mutation’s intrinsic fitness effect relative to the quality of the genetic background on which it occurs. Long-term evolutionary dynamics in these populations are therefore driven primarily by the stochastic introduction of mutants at the high-fitness tip of the population’s fitness distribution and the fluctuation in the lineage sizes of these super-fit mutants when rare. Several models have been introduced to study evolution in these strong selection, strong

Copyright © 2013 by the Genetics Society of America

doi: 10.1534/genetics.113.156430

Manuscript received May 11, 2013; accepted for publication August 27, 2013

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.156430/-/DC1>.

¹Corresponding author: Departments of Organismic and Evolutionary Biology and of Physics, FAS Center for Systems Biology, Harvard University, 435 Northwest Labs, 52 Oxford Street, Cambridge, MA 02138. Email: mdesai@oeb.harvard.edu

mutation regimes (Tsimring *et al.* 1996; Rouzine *et al.* 2003; Desai and Fisher 2007; Hallatschek 2011). This work has successfully described the rate of adaptation and the variation in fitness within a population (Desai and Fisher 2007; Park *et al.* 2010; Rouzine *et al.* 2008), and the fitness effects of fixed mutations (Fogle *et al.* 2008; Neher *et al.* 2010; Good *et al.* 2012; Fisher 2013), while ignoring the specific mutations that underlie these population-wide quantities.

In this work, we use these earlier theoretical treatments as the basis for analyzing the evolutionary dynamics of individual mutations (*i.e.*, their frequencies over time and their eventual fates). To do so, we study the forward-time dynamics of specific mutant lineages on the backdrop of the population's fitness distribution, concentrating specifically on quantities relevant and measurable in experiments: the trajectories and sojourn times of polymorphic sites, fixation times of successful mutants, and some basic patterns of genetic diversity.

We begin by introducing our model and briefly summarizing earlier results that describe the dynamics of the population's fitness distribution. We then demonstrate that the growth of the high-fitness “nose” of this distribution is dominated by a small number of successful, founding mutants. Since this high-fitness nose will eventually come to dominate the population, the long-term success of a given polymorphism is largely determined by its representation (or lack thereof) among this small class of stochastically fluctuating, high-fitness individuals. This allows us to model adaptation as a series of replacements of each fittest class by a new, fitter class over a typical replacement timescale. We show how this leads to a distribution of transition probabilities that describe how the frequency of each polymorphism changes in each stochastic jump from one fittest class to the next. This process bears some resemblance to several recent models of adaptation in populations with highly skewed offspring distributions (Schweinsberg 2003; Eldon and Wakeley 2006; Der *et al.* 2012). However, whereas in these earlier models a jump in offspring frequency is assumed to be an explicit feature of the offspring distribution, in this work these jumps emerge organically from the dynamics of the underlying model.

We next use our derived transition probabilities to calculate various diversity statistics, providing an alternative forward-time perspective that complements earlier structured coalescent approaches to these questions (Desai *et al.* 2013). We first calculate the site frequency spectrum of beneficial and neutral mutations, which has not yet been explicitly derived for this class of models. We then use our results to make predictions regarding the fates of mutations in experiments, particularly on the sojourn time of these mutations and the time to fixation of a successful mutant. Finally, in the *Discussion* we describe a *decay to neutrality* exhibited by mutations in these populations, comment on the relationship between our results and the Bolthausen–Sznitman coalescent, analyze the implications of our results for interpreting widely used tests for adaptation, and consider the extension of our model to evolution on more complex fitness landscapes.

Our work complements earlier analysis of related questions in facultatively sexual populations (Neher and Shraiman 2011), which neglects new mutations and focus instead on fluctuations in the frequencies of individual polymorphisms driven by recombination into higher or lower fitness backgrounds. By contrast, we focus on either asexual populations or on tightly linked genomic regions of sexual populations, where recombination can be neglected compared to selection and new mutations, and study instead the fluctuations in polymorphism frequencies driven by new mutations. Despite the different sources of draft explored in these two models, our results bear some similarities to the findings of Neher and Shraiman (2011), a parallel that we revisit in *Discussion*.

Model

We study the evolution of a large asexually reproducing population of constant size N , using the model introduced in Desai and Fisher (2007) (summarized below). This model assumes that beneficial mutations of a single fitness effect s occur at a constant beneficial mutation rate U_b per genome and are drawn from an effectively infinite number of possible sites. The use of a single fitness effect allows the fitness of an individual in the population to be described solely by the number of beneficial mutations k it carries, with the absolute fitness of an individual given by $w_k = (1 + s)^k \approx 1 + ks$ for $s \ll 1$. More complicated effects, such as frequency-dependent selection and epistasis, are neglected in this analysis (although the model is easily modified to include some simple sign epistasis; see *Discussion*). Finally, we assume that the population is evolving in the strong selection, strong mutation regime. Specifically, this means that $Ns \gg 1$, $s/U_b \gg 1$, $NU_b \gg 1$, meaning that the selective forces, selective forces relative to mutations, and incoming mutations per generation are all large.

In the next few paragraphs we review the primary features of this model that are pertinent to our analysis, which are justified in detail by Desai and Fisher (2007). This model describes the population as a traveling wave in fitness space, wherein the deterministic evolution in the bulk of the wave is combined with a careful stochastic treatment of the birth and fluctuation in lineage sizes of mutants at the high-fitness nose of the distribution. Specifically, the population is characterized according to the number of individuals n_k in each fitness class k , where the term *fitness class* refers to the class of individuals carrying k beneficial mutations. At each generation, n_k changes according to the effects of genetic drift, incoming and outgoing mutations, and selection. If n_k is sufficiently large, then selection trumps the effects of the other two evolutionary forces, and the rate of change of n_k is

$$\frac{dn_k(t)}{dt} \approx n_k(ks - \langle ks \rangle), \quad (1)$$

where t is taken in units of generations and $\langle ks \rangle$ is the (time-dependent) mean fitness of the population. A fitness class

will enter this regime of deterministic growth shortly after the effect of selective forces overcomes the effect of drift, which occurs shortly after the entire fitness class reaches a population size $\sim 1/(k - \langle k \rangle)s$, at which point we say that it is *established*. Given our assumption that $s \gg U_b$, the probability for a fitness class that has not yet established to generate a more fit establishing lineage is extremely low. Thus, the population is well described by a deterministically growing/shrinking set of fitness classes $\{n_{k_{\min}}, n_{k_{\min}+1}, \dots, n_{k_{\max}-1}\}$ and one stochastically fluctuating class $\{n_{k_{\max}}\}$, where k_{\min}, k_{\max} are defined to be the minimum/maximum k s.t. $n_{k_{\min}}, n_{k_{\max}} \neq 0$.

Although in principle one could consider the transient dynamics by which an initially clonal population attains a steady distribution of relative fitnesses, we are instead interested in the regime where this equilibrium distribution has already been reached and is maintained over timescales that are long compared to the typical establishment time of a new fitness class. In other words, the population has been evolving long enough to attain some typical steady-state fitness profile, but not long enough to begin to deplete the supply of beneficial mutations (which validates our infinite-sites approximation). In this case, the width of the distribution is set by an equilibrium between the influx and growth of highly fit mutants (which increases the width of the distribution) and the advancement of the mean fitness (decreasing this width). The size of this width q (defined as the mean number of mutations between the mean fitness class and the largest not-yet-established class) is given to a good approximation by

$$q \approx \frac{2 \log(Ns)}{\log(s/U_b)} \quad (2)$$

(Desai and Fisher 2007, Equation 4), with higher-order corrections given in Desai and Fisher (2007, Equation 39). Similarly, τ_k , defined roughly as the random variable denoting the time between the establishment of fitness class $k - 1$ and k , has expectation value

$$\langle \tau_k \rangle = \frac{1}{(q-1)s} \log \left[\frac{s}{U_b} \frac{q \sin(\pi/q)}{\pi e^{\gamma_E/q}} \right] \quad (3)$$

(Desai and Fisher 2007, Equation 36), with $\gamma_E \approx 0.577$ the Euler gamma constant. Note that the variable is called τ_q in Desai and Fisher (2007), where it contains an erroneous factor of $(q - 1)$ instead of q within the logarithm.

A more accurate derivation of the true time between establishments, accounting primarily for the non-negligible effect of incoming mutations shortly after a class establishes, is derived in Brunet *et al.* (2008), whereas a more careful discussion of the correct interpretation of τ_k is given in Desai and Fisher (2007). However, the precise distribution of τ_k is not important for our analysis because throughout the rest of this article we take time in units of establishment of each new lead class.

If q is not small, the mean relative fitness of any individual class does not change too much in the course of one establishment time. In this case, the dynamics of fitness

classes are excellently approximated by a *staircase model*, in which the fitness of every class is held constant over the course of the establishment of a new class. In the time it takes each new class to establish, the mean fitness will typically increase by s , and hence the relative fitness of all classes decreases accordingly. Thus, the number of individuals in each class grows exponentially at a rate that is well approximated as diminishing by discrete steps of s after each establishment time τ_k . Under these assumptions, the number of individuals in a class k with ℓ mutations more or less than the mean is approximately

$$n_k \approx \frac{1}{qs} \prod_{i=\ell}^{q-1} \exp(is \langle \tau_k \rangle) \approx \frac{1}{qs} e^{s \langle \tau_k \rangle (q(q-1) - \ell(\ell-1))/2}. \quad (4)$$

A more detailed discussion of this formula is given in Desai and Fisher (2007, p. 1774). Note that by averaging the fitness distribution over all times we recover a Gaussian distribution with variance $\sigma^2 = v = s/\langle \tau_k \rangle$, where v is the rate of adaptation.

We are primarily interested in the growth of fitness classes when they are still expanding near the high-fitness tip of the wave, which is where mutations destined to reach appreciable frequencies first occur. Thus, we examine the growth of the class $k = k_{\max}$, *i.e.*, the largest not-yet-established class. The number of individuals in this lead class at short times is (Desai and Fisher 2007, Equation 31)

$$n_k(t) = \frac{e^{qs(t-\tau_k)}}{qs}. \quad (5)$$

Note that we have arbitrarily defined the origin of time such that $t = \tau_k$ corresponds to the establishment time of class k , which occurs when $n_k = 1/qs$ individuals. By a simple transformation,

$$n_k(t) = \frac{\sigma_k e^{qs(t-\langle \tau_k \rangle) - \gamma_E/(q-1)}}{qs}, \quad (6)$$

where we have defined

$$\sigma_k = e^{-qs(\tau_k - \langle \tau_k \rangle) + \gamma_E/(q-1)}. \quad (7)$$

In formulation (5), the stochasticity of the growth is encapsulated in the establishment time τ_k ; in formulation (6), the random variable σ_k encodes how much the class deviates from its typical size at long times. The random variable σ_k is chosen to have the simple generating function [obtained by a transformation of the generating function of $n_k(t)$ given in Equation 29 of Desai and Fisher (2007)],

$$\langle e^{-z\sigma_k} \rangle = e^{-z^{1-1/q}}. \quad (8)$$

Note that σ_k is singularly well suited for extracting the contribution of independent lineages. To see this, we note that $n_k(t)$ denotes the growth of class k , given that it is fed by an exponentially (and deterministically) expanding class $k - 1$ (which we simply call the *feeding class*)

$$n_{k-1}(t) = \frac{1}{qs} e^{(q-1)st}, \quad (9)$$

which is supplying mutants to class k of fitness qs at a rate U_b per genome per generation. To probe the contribution of particular haplotypes from the feeding class, we decompose the growth of class $(k - 1)$ into

$$n_{k-1}(t) = \frac{\sum_{\ell} x_{\ell,k-1}}{qs} e^{(q-1)st}, \quad \sum_{\ell} x_{\ell,k-1} = 1, \quad (10)$$

where $x_{\ell,k-1}$ is the fraction of the feeding class constituted by haplotype ℓ . Note that if each haplotype is plentiful enough to expand deterministically (an assumption we discuss in more detail below), then each haplotype grows at the same rate as the class as a whole, and the $x_{\ell,k-1}$ are time-independent constants. In this case, the growth of the lead class may be written as

$$n_k(t) = \frac{\sigma_k}{qs} e^{qs(t-(\tau_k))-\gamma_E/(q-1)} = \frac{\sum_{\ell} \nu_{\ell,k}}{qs} e^{qs(t-(\tau_k))-\gamma_E/(q-1)}, \quad (11)$$

where $\nu_{\ell,k}$ is the random variable denoting the contribution from haplotype ℓ to class k from class $k - 1$. Note that it is *not* necessary that $\sigma_k = \sum_{\ell} \nu_{\ell,k} = 1$; in fact, σ_k may deviate strongly from unity if a mutation occurs and establishes in the lead class anomalously early. As a result, $\nu_{\ell,k}$ is not the frequency of the lineage derived from haplotype ℓ in class k . Rather, $\nu_{\ell,k}$ is the random variable that encodes how quickly that lineage establishes and expands in class k relative to its typical growth. The frequency of the derived lineage, then, is $x_{\ell,k} = \nu_{\ell,k}/\sigma_k$. This setup is illustrated in Figure 1. Since each haplotype in class $k - 1$ expands exponentially, the lead class k is fed by m independent feeding processes, where m is the total number of haplotypes in class $(k - 1)$. From this, we can essentially repeat the derivation for the generating function of σ_k done in Desai and Fisher (2007) for $\nu_{\ell,k}$ by adding appropriate factors of $x_{\ell,k-1}$ in the growth of the feeding class, eventually obtaining

$$\left\langle e^{-z\nu_{\ell,k}} \right\rangle = e^{-x_{\ell,k-1}z^{1-1/q}}. \quad (12)$$

Note that the assumption of independence between each $\nu_{\ell,k}$ implicitly assumes that the size of the lead $n_k \ll N$. In this case, feedback effects, whereby the growth of each $\nu_{\ell,k}$ affects the advancement of the mean fitness, and thereby the growth of other $\nu_{\ell,k}$, may safely be neglected. When class k leaves the lead and this assumption begins to break down, the frequency of each lineage ℓ derived from the corresponding haplotype in class $k - 1$ is already frozen in class k (an assumption justified in detail in the *Appendix*) and the lineage is expanding deterministically. Furthermore, because the rate of expansion of both the lineage and the class as a whole are set by the relative fitness of the class, at long times the frequencies $x_{\ell,k}$ do not change. The process will

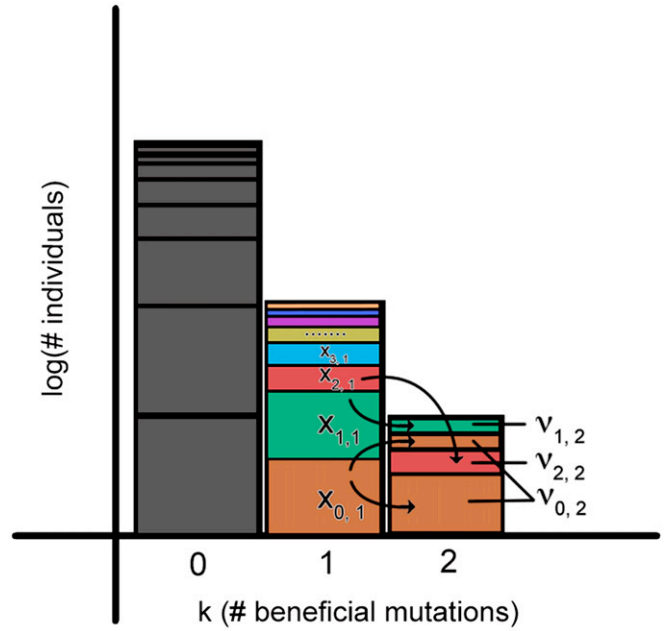


Figure 1 Illustration of the stochastic process by which lineages in a class $k - 1$ contribute to a class k through further beneficial mutations. Shown are the three fittest classes of the traveling wave. Class 1 is broken down into haplotypes $\{0, 1, 2, \dots, \ell\}$ at frequencies $\{x_{0,1}, x_{1,1}, \dots, x_{\ell,1}\}$ within class 1, which grow deterministically. At the instant illustrated here, haplotype 0 (at frequency $x_{0,1}$ within class 1) has generated two beneficial mutations. These mutations occurred at two different timepoints in the past, grew stochastically, and now cumulatively form a significant fraction of class 2. Individuals derived from haplotype 0 now form a fraction $x_{0,2} = \nu_{0,2}/(\nu_{0,2} + \nu_{1,2} + \nu_{2,2})$ of this class. Analogous observations can be made for the other haplotypes. Before class 2 begins generating mutants destined to establish in class 3, the frequencies of its haplotypes will be frozen up to small fluctuations, and the process repeats with a new set of frequencies $\{x_{0,2}, x_{1,2}, \dots, x_{\ell,2}\}$.

then begin anew, with a new set of haplotypes in class k at frequencies $x_{\ell,k}$ feeding a new most-fit class, $k + 1$. Hence, the number of individuals descended from a lineage ℓ at a time t , $\eta(t)$, is given simply by $\eta(t) = \sum_{k=k_{\min}}^{k_{\max}} x_{\ell,k} n_k(t)$.

Some objections to this formalism might immediately be raised. First, the number of haplotypes m in class $k - 1$ is increasing in time due to incoming mutations from the previous class $k - 2$. However, we show in the *Appendix* that incoming mutations typically stop contributing significantly to a class shortly after it establishes, and certainly by the time that it itself begins feeding establishing mutants to the next class. Thus, while m may be strictly increasing in time, the combined contribution of these new haplotypes affects the frequencies of all other sites only negligibly. Second, a considerable fraction of these haplotypes could be at frequencies $x_{\ell,k-1}$ such that $x_{\ell,k-1} n_{k-1}(t) \lesssim 1/((q - 1)s)$, meaning that these haplotypes cannot be modeled deterministically. This objection becomes important when considering fluctuations in the frequencies of haplotypes that are rare in a given class. On the other hand, if a polymorphic site $x_{\ell,k-1}$ is sufficiently common, its growth by the time the class begins supplying establishing mutants may be modeled

deterministically. In this case, the above formalism is well suited to predict the contribution of that lineage ℓ to subsequent fitness classes.

In what follows, since we are interested only in the dynamics of one particular lineage at a time, we drop the ℓ subscript and set the initial fitness class (*i.e.*, the fitness class in which the lineage frequency first begins to be tracked) at $k = 0$. The strengths and realm of validity for all of the above assumptions have been studied by ourselves and others in previous work (Desai and Fisher 2007; Brunet *et al.* 2008; Rouzine *et al.* 2008; Desai *et al.* 2013; Fisher 2013).

Transition Probabilities

From the considerations of the previous section, we see that evolutionary dynamics in these populations are driven by two factors: the deterministic growth and decay of existing clones—governing the short-time dynamics—and the stochastic introduction and expansion of new super-fit mutants, which govern the population’s long-term evolution. Furthermore, the fate of any particular mutation is determined by two factors: (1) the genetic background in which it occurs and (2) the success of its progeny in amassing additional beneficial mutations more quickly than competing backgrounds. When many beneficial mutations segregate simultaneously, the strictness of these two constraints requires that mutations with any nonnegligible chance of fixing (or even rising to an appreciable frequency) must have been founded near the nose of the fitness distribution. The vast majority of beneficial mutations are thus “wasted” on the bulk of the distribution, where the mutant’s lineage is doomed to eventual extinction.

We are largely concerned with those mutations that successfully rise to appreciable frequencies and thus largely consider the dynamics of those mutations that are founded on “good” genetic backgrounds. Once established, one of these new, high-fitness clones will grow, stagnate, and diminish in its class deterministically as dictated by Equation 1. At any moment in time, the population can be divided into many such expanding and contracting “bubbles,” which fully determine frequency dynamics over short timescales of $\mathcal{O}(\tau_k)$. However, many of the interesting long-term dynamics are determined by the stochastic origination and establishment of super-fit mutants from these deterministically expanding clones, which drive the success or failure of particular lineages, mutations, or entire evolutionary trajectories.

These ideas are expressed more concretely in Figure 2, which shows the distribution of fitness classes at three distinct timepoints. A clone that is about to establish in the first timepoint is growing deterministically in the second timepoint and diminishing in the third timepoint, before finally going extinct as the population evolves to higher and higher fitness. Normally, this would mean that the contribution of the clone’s lineage is also extinct; however, the lineage avoids this fate by jumping into the next fitness class through the creation of a new, super-fit mutant when the class is still small and expanding very rapidly. This new mutant establishes and

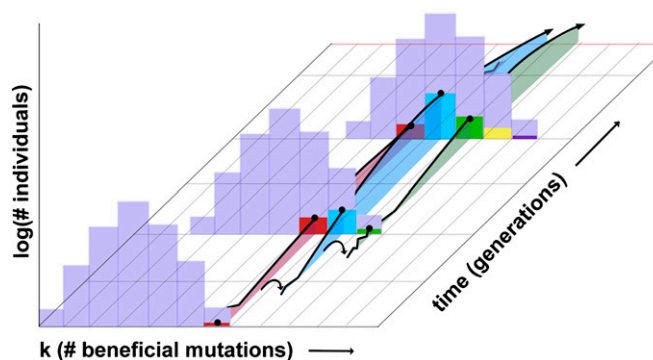


Figure 2 The frequency trajectory of a successful mutation on a high-fitness background. The population’s fitness profile is shown for three timepoints, along with one haplotype present in each class. A mutation founded in the first timepoint (in red) expands and constitutes a significant fraction of its fitness class in the second timepoint. As the population adapts, the founding fitness class begins to shrink, along with the number of individuals in the lineage constituting that class. However, this lineage generates a beneficial mutant (shaded in blue) between the first and second timepoints, as shown by the first arrow. This mutant drifts randomly for a short while before establishing and eventually comprising a significant fraction of the next fitness class. In this way, the original successful mutation “jumps” into the next fitness class and is potentially able to avoid extinction. This process repeats, as the second (blue) lineage generates another successful (green) mutant that comes to populate a significant fraction of the next class, and so on.

expands, and because it occurs very early, comes to form a significant fraction of the next fitness class, as exemplified in the second timepoint of Figure 2. Although only one of these jumps is shown in Figure 2, a given lineage may jump many times into the next class, with each successive jump, on average, contributing a smaller and smaller fraction of individuals to that class. This new clone will then deterministically expand, contract, and go extinct in the new class, although its lineage may survive by jumping into the next fitness class sufficiently early to eventually constitute a significant fraction of that class, shown by the second jump in Figure 2. The process continues *ad infinitum*, until the sum of all the contributions of a given lineage to a class vanishes or constitutes the entire class, in which case the lineage is then destined to go extinct or sweep, respectively.

The key distribution describing these dynamics is the jump probability $\rho(x_k|x_{k-1}, \dots, x_0)$, the probability of finding a lineage at frequency x_k in fitness class k , given that it was at frequencies $x_{k-1}, x_{k-2}, \dots, x_0$ in the k previous fitness classes. Essentially, $\rho(x_k|x_{k-1}, \dots, x_0)$ gives the probability distribution of the sum of the frequencies of each clone in fitness class k that originated from a lineage at some frequency x_0 in fitness class 0 and jumped through $k - 1$ intermediate classes. Under our particular model, the derivation of $\rho(x_k|x_{k-1}, \dots, x_0)$ becomes much simpler because the frequency of a lineage in fitness class k is determined only by its frequency in fitness class $k - 1$. This is equivalent to the statement that when a class begins feeding establishing mutants to the next class (*i.e.*, when the jumps in Figure 2 occur), the frequencies of lineages in the feeding class are already frozen. In this case,

one may consider the frequencies of lineages evolving in analogy to the entire population: the frequencies of lineages in the classes $\{n_{k_{\min}}, n_{k_{\min}+1}, \dots, n_{k_{\max}-1}\}$ are frozen, and the frequencies of mutants in the lead $\{n_{k_{\max}}\}$ are fluctuating. Thus the transition process is a Markov chain, meaning that the long-time frequency of a lineage in fitness class k , x_k , is simply $\rho(x_k|x_{k-1}, \dots, x_0) = \rho(x_k|x_{k-1})$. Of course this viewpoint sacrifices precision for the sake of clarity, since frequencies of lineages in the lead will continue to fluctuate after the lead establishes. It is only when a class typically begins to supply *establishing* mutants to the nose that the frequencies of its lineages will be frozen. Regardless, there will typically be one class with fluctuating frequencies (either the lead or next-to-lead class shortly after it establishes) and the rest of the population with lineage frequencies already frozen. This is by no means an obvious assumption and is discussed in detail in the *Appendix*.

In Equation 12, we gave the generating function of ν_k (the contribution of a particular lineage to a fitness class

k , given that the frequency of the lineage in class $k-1$ is x_{k-1}). In class k , the lineage will grow as

$$\eta_k(t) = \frac{x_k}{qs} e^{(q-1)s(t-\tau_k)} \propto \frac{\nu_k}{qs} e^{(q-1)s(t-\langle\tau_k\rangle)}, \quad (13)$$

where the exact proportionality constant is not relevant for our analysis. Similarly, we may denote the contribution of that lineage's complement by $\tilde{\nu}_k$ —that is, the contribution to k from those individuals whose ancestors in class $k-1$ were *not* derived from the chosen lineage. Naturally, $\tilde{\nu}_k$ is described by the following generating function,

$$\langle e^{-\tilde{\nu}_k z} \rangle = e^{-(1-x_{k-1})z^\alpha}, \quad (14)$$

with $\nu_k + \tilde{\nu}_k = \sigma_k$, $\alpha = 1 - 1/q$ and σ_k defined in Equation 6. Then, $\rho(x_k|x_{k-1})$ is derived from the two generating functions to be

$$\rho(x_k|x_{k-1}) = \frac{\sin(\pi\alpha)x_{k-1}(1-x_{k-1})}{x_k(1-x_k)\pi \left[(1-x_{k-1})^2(x_k/[1-x_k])^\alpha + x_{k-1}^2([1-x_k]/x_k)^\alpha + 2x_{k-1}(1-x_{k-1})\cos(\pi\alpha) \right]}, \quad (15)$$

which is corroborated by results of forward-time simulations (see Figure 3). This is readily extended to $\rho(x_k|x_0)$, the distribution of a given lineage frequency k fitness class steps forward, by the simple replacement $\alpha \rightarrow \alpha^k$. The derivation of Equation 15 and the extension to arbitrary time steps k are given in the *Appendix*. Note that this transition probability was also stated, without detailed derivation, in Desai *et al.* (2013). Importantly, parameters such as N , s , and U_b factor into this distribution only through the parameter $\alpha = 1 - 1/q \approx 1 - \log(s/U_b)/(2 \log(Ns))$.

Note that while the result for one time step is exact, the generalization to several steps forward assumes that $\sigma_{k'} \approx 1$ for $0 < k' < k$, or that the rate of establishment of each new class does not deviate too much from typical values. If this is the case, then $\nu_{k'} \approx x_{k'}$, and each lineage may be assumed to grow independently of its competition in each new establishment. Errors introduced using this method start to become significant for $k \approx q$, which is the timescale at which fluctuations in the establishment time of new lead classes begin to affect the rate of adaptation and mean fitness. We later demonstrate that calculations using $\rho(x_k|x_0)$ yield robust estimates for a variety of diversity statistics and other population-wide properties, supporting our claim that the above assumption does not significantly change most quantitative predictions.

Although the analytic form of this jump probability is somewhat cryptic, some illuminating properties can be gleaned by considering the form of the distribution in the limit of infinite population size, $q \rightarrow \infty$:

$$\rho_{q \rightarrow \infty}(x_k|x_{k-1}) = \frac{x_{k-1}(1-x_{k-1})}{q(x_{k-1}-x_k)^2}. \quad (16)$$

Because this distribution has a fat tail, large jumps may occasionally occur between adjacent fitness classes even in the limit of a massively large population, indicating that the dynamics described by this stochastic process are superdiffusive. This nondiffusive property of genetic draft has been observed many times for a number of different models of adaptation (Gillespie 2001; Neher and Shraiman 2011; Desai *et al.* 2013; Neher and Hallatschek 2013). One important, commonly stressed consequence of this nondiffusivity is that the stochastic dynamics of the population cannot be fully encapsulated by a single rescaled variance effective population size N_e .

Implications for Genetic Diversity

Whereas ρ is not directly measurable by itself—describing transition probabilities between fitness classes and not the population as a whole—genetic draft leaves distinct signatures on the diversity of rapidly adapting populations, which are both readily measurable and readily derived from ρ . In what follows we derive some implications of the stochastic jump process on the site frequency spectra of both beneficial and neutral mutations.

The site frequency spectrum of beneficial mutations

One statistic that is strongly affected by the stochastic jumps described above is the site frequency spectrum (SFS) of beneficial mutations $f(x)$, the expected density of mutations between frequencies x and $x + dx$. In rapidly adapting populations, the SFS is partitioned into two regimes: on one hand, common mutations first arise at the distribution's high fitness nose and are strongly affected by the process of

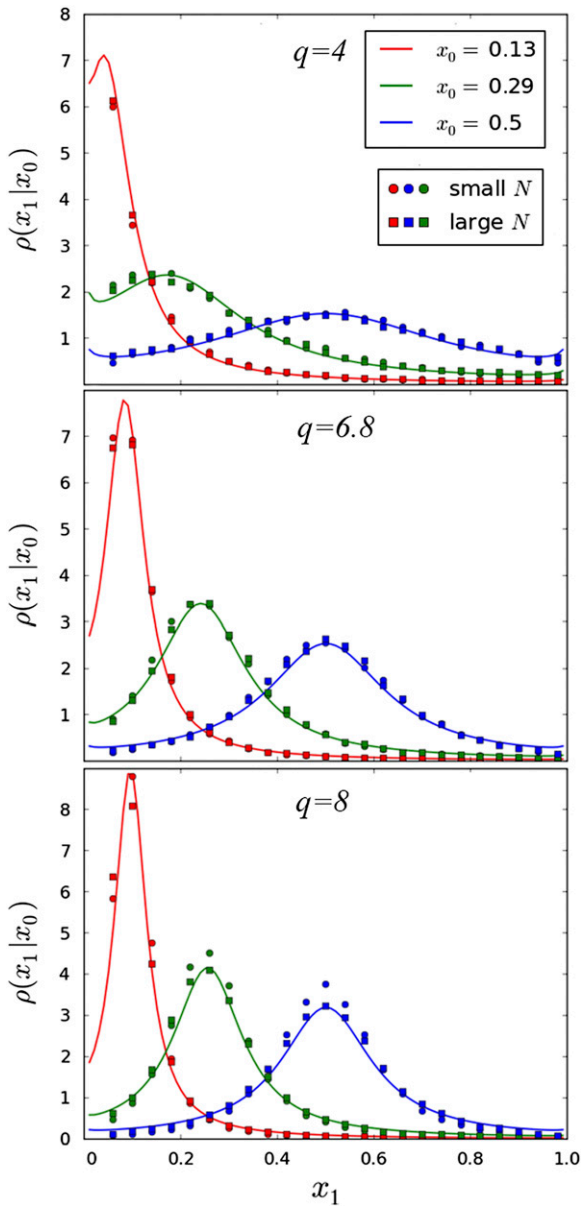


Figure 3 The transition probability $\rho(x_1|x_0)$ plotted for different values of x_0 , q , and other population parameters U_b , s , and N . Curves represent the theoretical prediction for the transition probabilities given in Equation 15. Circles and squares represent the results of forward-time simulations. The circular and square points have identical values of q , but differ in population size by a factor of 10. Circles (small N) represent simulations run with parameters $s = 0.1$, $U_b = 0.001$, and squares (large N) are run with parameters $s = 0.01$, $U_b = 0.0001$. The three plots represent different values of q , with $q = 4$ calculated from the approximate value in the text, and $q = 6.8, 8$ calculated with higher-order corrections given in Desai and Fisher (2007). These corrections become more important for higher q and are the reason for the slight discrepancy between theorized and predicted values for small N at $q = 8$. More information about the simulations may be found in File S1.

stochastic jumps. On the other hand, nearly private variants, which constitute the majority of beneficial mutations, are overwhelmingly founded near the distribution's bulk and are largely unaffected by this process. Thus, we expect the

high- and low-frequency spectra to be qualitatively different. As a result, the derivation that follows is split into two segments: first, we derive the SFS of common alleles, which are founded at the exponentially expanding wavefront. Since our previous analysis describes only the dynamics of these exponentially expanding nose classes, it does not describe the frequency spectrum of extremely rare, nearly private alleles. Thus, in the second part of this section we make use of a different branching process method to derive the distribution of these plentiful but extremely rare variants.

We begin by deriving the site frequency spectrum of common alleles. For the moment, we may further simplify the problem by considering the site frequency spectrum of mutants in only one fitness class. Since the frequencies of common mutations freeze shortly after a given class establishes, we need only to calculate the distribution of frequencies in a class that is near the nose of the wave. Once the class leaves the nose, the frequencies of these common mutants will be frozen and the class will have the same SFS regardless of its position relative to other classes.

Now, as we have previously argued, almost every common polymorphism was once founded in a class that was at the population's high-fitness nose. Thus, given a fitness class k that is near the distribution's nose, we might decompose the frequency spectrum of mutants in class k according to which class they were founded in. Specifically, we consider mutations founded in class k from class $k - 1$ when class k was at the nose; these mutations originate in class k . Next, we may consider mutations that originated in class $k - 1$ from class $k - 2$ when class $k - 1$ was near the nose, whose lineages subsequently jumped into class k after acquiring more beneficial mutations; these mutations originate in class $k - 1$. Analogously, we can consider the distribution of mutants in class k that originated in classes $k - 2$, when these classes were at the distribution's nose. The SFS $f(x)$ in class k is then the sum over all of these distributions.

Consider first the SFS of sites in class k that originate in k , $f_1(x)$. We derive the SFS of these "new" mutations from the transition probability ρ as follows: first, we observe that class $k - 1$ first begins supplying mutants to class k that are *destined to establish* at a time of $\mathcal{O}(\langle\tau_k\rangle)$ since its own establishment. Shortly before this time, we can decompose the growth of class $k - 1$ into $1/\beta$ independently growing blocks of frequency β such that $\beta \ll 1/q$. If β is sufficiently large, so that $\beta n_{k-1}(\langle\tau_k\rangle) \gg 1/qs$, it is valid to model the growth of each block deterministically. In this case, the frequency x in class k of descendants of individuals in block β is distributed as $\rho(x|\beta)$. Since $n_{k-1}(\langle\tau_k\rangle) = U_b^{-1}$, we require that $1/q \gg \beta \gg U_b/qs$ for the jumps from each β -sized block into class k to be well described by $\rho(x|\beta)$. Now, since $\beta \ll 1/q$, it is unlikely that more than two founding mutants originate from the same block. In other words, the contribution of each block to the next class is dominated by the contribution of one most successful mutation, and the probability distribution of the frequency x of this most successful mutation is well approximated by $\rho(x|\beta)$. The expected

density of mutations between frequencies x and $x + dx$ that are introduced from class $k - 1$ is then

$$f_1(x) = \frac{\rho(x|\beta)}{\beta} \approx \frac{\sin(\pi\alpha)(1-x)^{\alpha-1}}{\pi x^{1+\alpha}}. \quad (17)$$

Analogously, the distribution of mutations in class k originally arising from the $(k - i)$ th class is obtained by the replacement $\alpha \rightarrow \alpha^i$, giving for the total SFS

$$\begin{aligned} f(x) &= \sum_{i=1}^{\infty} \frac{\sin(\pi\alpha^i)(1-x)^{\alpha^i-1}}{\pi x^{1+\alpha^i}} \\ &\approx \frac{1}{\pi x(1-x)} \int_0^{\infty} dt \sin(\pi\alpha^t) \left(\frac{1-x}{x}\right)^{\alpha^t} \\ &= \frac{1}{\pi x(1-x)\log(q/(q-1))} \int_0^1 dz \frac{\sin(\pi z)}{z} \left(\frac{1-x}{x}\right)^z. \end{aligned} \quad (18)$$

Although straightforward to evaluate numerically, this integral has no simple closed-form expression. However, a first-order Taylor expansion in the sine is a reasonable approximation for $1 - x \ll 1$. This gives

$$f(x) \approx \frac{1 - 2x}{\log(q/(q-1))\log((1-x)/x)x^2(1-x)}, \quad 1 - x \ll 1. \quad (19)$$

For $x \rightarrow 1$,

$$f \sim ((1-x)\log(1-x))^{-1}. \quad (20)$$

Note that this predicts that very-high-frequency mutations are actually more common than mutations at slightly lower frequencies. This upswing at very-high-frequency mutations is a widely recognized marker of selection for a number of different models, with and without linkage between sites (Wright 1938; Fay and Wu 2000; McVean and Charlesworth 2000; Neher and Hallatschek 2013). Importantly, it cannot be explained by many commonly studied forms of demographic history, such as population expansions. Recently, Neher and Hallatschek (2013) derived the upswing of the SFS arising from genealogies obeying the Bolthausen–Sznitman coalescent. However, to our knowledge, the excess of extremely common variants arising from linked beneficial mutations has not previously been derived directly from the forward-time dynamics of any model.

Another point worth mentioning is the *invariance* of the functional form of this distribution for different parameters N , s , and U_b . Given that our assumptions about the population hold, the spectra of different populations are *identical* up to a scaling factor that represents the different absolute numbers of common mutations in these populations, which itself is somewhat insensitive to the specific choice of parameters. This observation highlights an inherent limitation in inferring properties of adaptation through the functional form of the site frequency spectrum.

We now argue that the above distribution is a good approximation to the population-wide site frequency spectrum, instead of simply the frequency spectrum of mutations in a single class. First, we observe that we can arbitrarily set the distribution $f(x)$ to describe the SFS of the mean class. The above approximation, then, is equivalent to the statement that the SFS of the mean class is a good approximation to the SFS of the entire population. If q is not too large (the relevant case for many biological populations), then the approximation holds because the vast majority of individuals reside in the mean class at any given time. The contribution of sites from other classes will then be a small perturbation on the SFS of the mean class, particularly relevant at low frequencies (where our approximation breaks down regardless, due to the contribution of mutants not founded at the nose). On the other hand, as q increases, the mean class constitutes a smaller and smaller fraction of the total population. However, the variance of the jumps in the frequencies of mutant sites also decreases in proportion to $1/q$. Thus, while the mean class constitutes a smaller fraction of the total population, sites in adjacent classes tend to shift more slowly than for the case of smaller q (despite the fact that, as we have previously noted, large jumps may still occur occasionally). Thus the approximation should still be valid even in the limit that q is large. The strength of these arguments is corroborated by the close correlation of Equation 18 with site frequency spectra derived from our forward-time simulations (Figure 4).

As we have previously mentioned, this distribution describes only frequencies of mutations that are founded in the exponentially expanding, high-fitness front of the wave. As such, it fails for rare mutations, which are overwhelmingly dominated by mutations that are introduced when the class is near the mean of the distribution. A different approach is then necessary for understanding the spectrum of these extremely low-frequency mutations.

Fortunately, all the difficulties in accounting for effects of genetic draft and the stochasticity of the wavefront are no longer a factor when dealing with these rare variants. By definition, the lineage of a rare variant never comprises a substantial fraction of the population and only rarely acquires further (establishing) beneficial mutations. Thus, the site frequency spectrum of these individuals can be studied using standard branching process methods given some fixed death rate $d = 1$ and a diminishing birth rate $b(t) = 1 + y_0 - vt$, where t is the time in generations, y_0 is the initial (relative) fitness of the mutant, and v is the mean rate of adaptation of the population. Furthermore, because these mutants occur in the bulk of the fitness distribution, where the number of individuals is very large, we can assume that they occur deterministically at some rate $U_b n_{k-1}(t)$.

In supporting information, [File S1](#), we show that as a consequence of these assumptions, the expected number of mutations with lineage sizes n in a class at fitness $y = 0$, $F_{\text{rare}}(n/N, 0)$, is given by

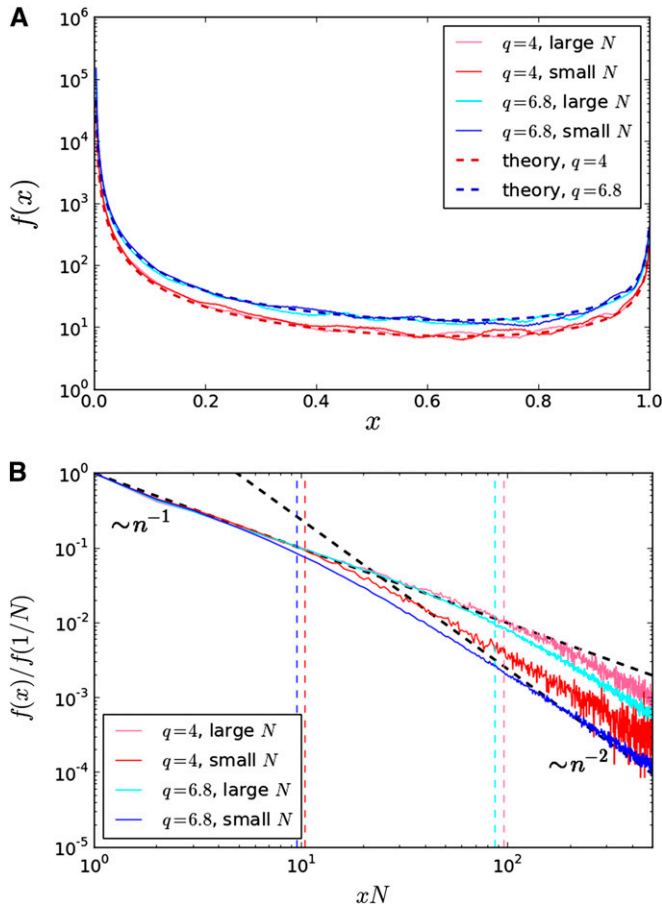


Figure 4 (A) The site frequency spectrum of common mutations. The dashed lines are the theoretical predictions from Equation 18 for the two tested values of q . Colored lines are frequency spectra derived from the results of 1000 forward-time simulations for each parameter set. For each evolved population, 1000 individuals were sampled from the final timepoint of an initially clonal population evolved for 15000 generations. s and U_b for small and large population sizes are the same as those quoted in Figure 3. (B) The site frequency spectrum of beneficial semiprivate variants (a magnification of A in the regime $x \ll 1$). The site frequency spectrum for each parameter set is normalized by the number of purely private mutations (i.e., sites with lineage sizes $n = 1$). The $1/n$ dashed black line is the decay predicted from Equation 21 for small n , and the $1/n^2$ dashed black line indicates the crossover to the (approximately) $1/n^2$ decay predicted for more common alleles. Dashed vertical lines represent the predicted realm of validity of the $1/n$ decay for each parameter set, derived in File S1. Colored lines are frequency spectra derived from the results of 1000 forward-time simulations for each parameter set. For each evolved population, the total number of mutations carried by n individuals were explicitly tabulated from the final timepoint of an initially clonal population evolved for 15000 generations. Parameter sets are the same as those quoted in Figure 3.

$$F_{\text{rare}}(n/N, 0) \approx \frac{U_b N e^{-s^2/(2v)}}{n \sqrt{2\pi v}}, \quad (21)$$

under the constraints that $n \ll \sqrt{2/v} - s/(2v)$ and $s \ll \sqrt{v}$. Significantly, the density of sites at the rare end of the frequency spectrum, $f_{\text{rare}}(n/N) \propto F_{\text{rare}}(n/N, 0) \propto 1/n$. Note that in comparing to the Wright–Fisher results we must multiply Equation 21 by a factor of ~ 2 , which reflects the different

stochastic dynamics of the branching process and Wright–Fisher model.

It is important to observe that these frequencies are at the extremely low end of what is colloquially considered to be a “rare” variant, and hence we dub these mutations more precisely as “nearly private” or “semiprivate.” For the frequencies commonly measured in a reasonably sized population sample, rare but non-singleton variants will still decay as $1/x^2$, and the effect of this skew for nearly private mutations will manifest itself as a smaller number of singletons than that predicted simply by the $1/x^2$ extrapolation.

The frequencies of mutant sites thus fall into two regimes. Common alleles founded at the wavefront have distributions similar to those of exponentially expanding populations. Conversely, semiprivate variants are largely founded recently in the past and in the bulk of the distribution and as a result exhibit a neutral SFS. This latter property is only to be expected, since a mutant landing in the mean fitness class has neutral relative fitness by definition, regardless of the specific fitness effects of the mutations it carries. These findings are supported by our forward-time simulations, demonstrated in Figure 4.

There is necessarily some crossover region between the regime of more common alleles and semiprivate variants, occurring in the region of landing fitness $y_0 \sim (q - 2)s$. Near these fitnesses, newly founded sites are no longer well described as originating in the expanding, high-fitness front of the wave; however, over the course of their existence a sufficiently significant number of them may reach large enough lineage sizes to be affected by draft. In this case, the distribution of site frequencies in fitness classes near the population’s bulk is skewed by mutations that were not founded near the nose, but still jumped into fitter classes. Such lineages will certainly contribute a potentially nonnegligible number of sites at the rare end of the spectrum (at lineage sizes larger than the semiprivate ones we have studied). However, because they do not change the results qualitatively, they are neglected in this work. This choice is supported by the rapid crossover between the $1/x$ rare variant decay and the (approximately) $1/x^2$ decay predicted for more common alleles, as exemplified by Figure 4.

The site frequency spectrum of neutral mutations

Our method is similarly well suited for calculating the SFS of neutral mutations. Once a mutation is present at some frequency in a given class, its frequency in subsequent classes is *purely* determined by draft. Thus, the only difference between the beneficial and neutral cases is in the distribution of mutations first introduced in a given class when that class was at the distribution’s nose, $f_1(x)$. The effect of draft on these mutations i classes later is then obtained by convolution with the jump probability,

$$f_{\text{neut}, i > 1}(x) = \int_0^1 f_{\text{neut}, 1}(y) \rho_{i-1}(x|y) dy, \quad (22)$$

where $\rho_{i-1}(x|y)$ denotes the probability density of a mutation at frequency x in a class $i - 1$, given that it was at frequency y in class 0. The total site frequency spectrum is then obtained by summing over all timepoints, corresponding to all possible originating classes:

$$f_{\text{neut}}(x) = \sum_{i=1}^{\infty} f_{\text{neut},i}(x). \quad (23)$$

In File S1 we show that $f_{\text{neut},1}(x) \approx U_n/((q-1)sx^2)$ to leading order, where U_n is the per-genome neutral mutation rate. Similarly, $f_{\text{ben},1}(x)$, the distribution of newly introduced beneficial mutations in a given class, was derived in Equation 17 to be

$$f_{\text{ben},1}(x)dx = \frac{\sin(\pi\alpha)}{\pi} \frac{dx}{(1-x)^{1/q}x^{2-1/q}} \rightarrow \frac{dx}{qx^2} \quad (24)$$

for $q \rightarrow \infty$. In this limit, it is true that

$$\frac{f_{\text{neut},1}(x)}{f_{\text{ben},1}(x)} = \frac{U_n\pi}{\sin(\pi\alpha)(q-1)s} \approx \frac{U_n}{s}. \quad (25)$$

As a result,

$$\begin{aligned} f_{\text{neut},i}(x) &= \int f_{\text{neut},1}(y)\rho_{i-1}(x|y)dy \\ &\approx \int \frac{U_n\pi}{\sin(\pi\alpha)(q-1)s} f_{\text{ben},1}(y)\rho_{i-1}(x|y)dy \\ &= \frac{U_n\pi}{\sin(\pi\alpha)(q-1)s} f_{\text{ben},i}(x) \end{aligned} \quad (26)$$

and by extension (since this holds for each i),

$$f_{\text{neut}}(x) \approx \frac{U_n\pi}{\sin(\pi\alpha)(q-1)s} f_{\text{ben}}(x) \rightarrow \frac{U_n}{s} f_{\text{ben}}(x) \quad (27)$$

in the limit $q \rightarrow \infty$. Thus, in the limit of rapid adaptation, the neutral and beneficial site frequency spectra differ only by a scaling factor set by the rate of neutral mutation relative to the strength of selection. This relationship demonstrates the fact that, as beneficial mutations become more common, the relative importance of a mutation's intrinsic fitness effect diminishes relative to the quality of its genetic background. Thus, the relative site frequency spectra are roughly set by the rate at which neutral mutations accrue in the nose classes relative to beneficial mutations. Although only strictly true in the infinite adaptation limit, our simulations demonstrate that this approximation is already accurate for q as small as 4 (Figure 5).

To compute the site frequency spectrum of neutral semiprivate variants, the derivation follows identically as for beneficial mutations, with $s \rightarrow 0$, $U_b \rightarrow U_n$. The result may then immediately be written down to be

$$f_{\text{neut,rare}}(n/N) \propto F_{\text{neut,rare}}(n/N, 0) \approx \frac{U_n N}{n\sqrt{2\pi v}} \propto \frac{1}{n}, \quad (28)$$

which holds so long as $n \ll \sqrt{2/v}$. This $1/n$ dependence is demonstrated in Figure 5.

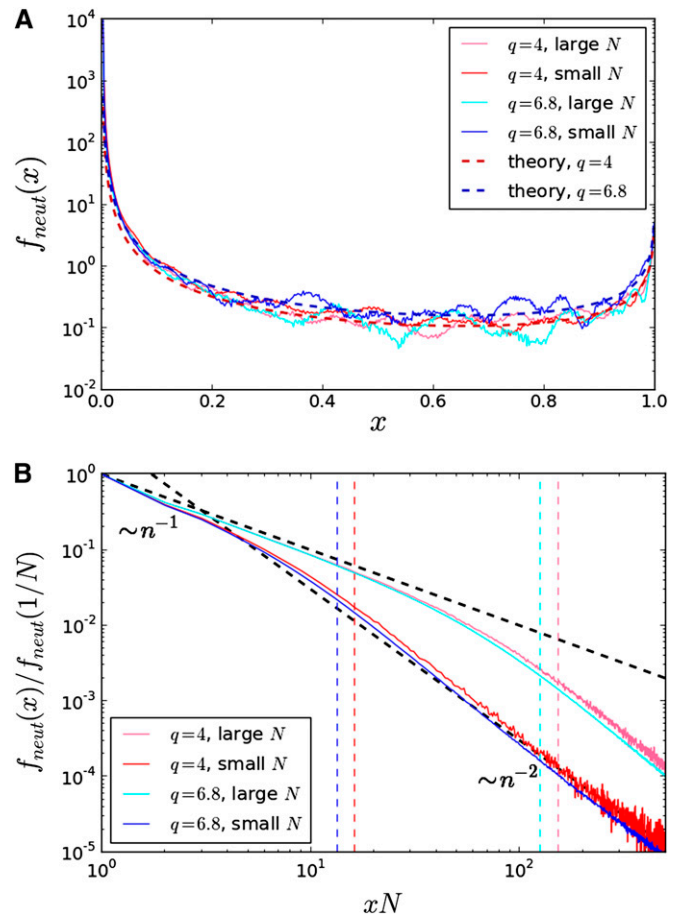


Figure 5 (A) The site frequency spectrum of common neutral mutations. The dashed lines are the theoretical predictions using the exact scaling in Equation 27 for the two tested values of q . Colored lines are frequency spectra derived from the results of forward-time simulations for each parameter set. For each evolved population, 1000 individuals were sampled from the final timepoint of an initially clonal population evolved for 15000 generations. Parameter sets are the same as those quoted in Figure 3, and additionally $U_n = U_b$ for each parameter set. (B) The site frequency spectrum of neutral semiprivate variants (a magnification of A in the regime $x \ll 1$). The site frequency spectrum for each parameter set is normalized by the number of purely private neutral mutations (*i.e.*, neutral sites with lineage sizes $n = 1$). The $1/n$ dashed black line is the decay predicted from Equation 28 for small n , and the $1/n^2$ dashed black line indicates the crossover to the (approximately) $1/n^2$ decay predicted for more common neutral alleles. Dashed vertical lines represent the predicted realm of validity of the $1/n$ decay for each parameter set, derived in File S1. Colored lines are frequency spectra derived from the results of forward-time simulations for each parameter set. For each evolved population, the total number of mutations carried by n individuals were explicitly tabulated from the final timepoint of an initially clonal population evolved for 15000 generations. Parameter sets are the same as those quoted in Figure 3, and additionally $U_n = U_b$ for each parameter set.

Sojourn Times

So far, we have described the effect of genetic draft on the frequencies of lineages as they jump through fitter and fitter fitness classes and characterized the effects of these jumps in skewing the resulting beneficial and neutral SFS. Now, we are ready to make predictions regarding fates and trajectories of observed polymorphic sites in these populations. One

of the most important predictions to be made is the time to fixation of a beneficial allele.

In contrast to the strong selection, weak mutation regime, in the strong selection, strong mutation regime, the lineage carrying a particular mutation usually jumps through many fitness classes before fixing in any one. Furthermore, it takes time for the class in which the mutation first fixed to traverse the length of the wave, adding $\sim 2q\langle\tau_k\rangle$ generations before the mutation is fixed in the population. Thus, in studying the fates of mutations, there are two pertinent questions. First, given a mutation at some measured frequency, how long does it take before the mutation sweeps or goes extinct? Second, given a newly established fitness class, how long does it take any mutation introduced in this class to sweep (equivalently, what is the expected time to fixation of a new mutation that is destined to fix)?

Because our method assumes that lineages in the feeding class are frozen once they begin to feed establishing mutants into the next class, our method is poorly equipped to deal with lineages at very high or low frequencies that are strongly affected by drift. Furthermore, the pathologies of our distribution (which, treating n_k as a continuous variable, allows for fractional numbers of individuals) introduces errors in the regime of $n_k \sim 1$. Nevertheless, we can calculate the sojourn time of these mutants by predicting when the frequency of a given mutation is expected to fall above or below a small threshold frequency ε . All the above problems may be circumvented if ε is taken to be small, but large enough for the lineage to be established in the feeding class when it begins supplying establishing mutants (roughly, this is fulfilled when $1 \gg \varepsilon \gg U_b/q_s$). When a lineage falls below frequency ε or rises above frequency $1 - \varepsilon$ in a given class, its probability of extinction or fixation is then $1 - \varepsilon$, which is nearly certain if ε is sufficiently small. We note that this scenario more accurately imitates what one could measure in an experimental setting, with a sample that is much smaller than the total population size or (if performing whole population sequencing) some finite-sequencing read depth. In these experimental scenarios, the absence of a particular polymorphism in such a measurement may not mean that the polymorphism is extinct entirely, but rather that it is unlikely to be present in the population above a certain frequency.

The distribution governing the sojourn time $P_{\text{soj}}(k|x_0)$ —the probability that a site at a frequency x_0 in fitness class 0 has a frequency x_k in fitness class k that freezes below the threshold ε (or above $1 - \varepsilon$)—is calculated as

$$\begin{aligned} P_{\text{soj}}(k|x_0) &= \int_0^\zeta \rho(x_k|x_0) dx_k + \int_{1-\varepsilon}^1 \rho(x_k|x_0) dx_k = \int_0^\zeta \Pi(\tilde{x}_k|x_0) d\tilde{x}_k + \int_{1/\varepsilon}^\infty \Pi(\tilde{x}_k|x_0) d\tilde{x}_k \\ &= \frac{x_0}{\pi} \text{Re} \left[\int_0^\zeta \frac{1}{-i(x_0\tilde{x}_k + (1-x_0)\tilde{x}_k^{1-k} e^{i\pi\alpha^k})} d\tilde{x}_k + \int_{1/\varepsilon}^\infty \frac{1}{-i(x_0\tilde{x}_k + (1-x_0)\tilde{x}_k^{1-k} e^{i\pi\alpha^k})} d\tilde{x}_k \right] \\ &= \frac{1}{\alpha^k \pi} \left[\arctan\left(\frac{x_0 \zeta^{\alpha^k} \sin(\pi\alpha^k)}{1 - x_0 \zeta^{\alpha^k} \cos(\pi\alpha^k)}\right) + \arctan\left(\frac{(1-x_0)\zeta^{\alpha^k} \sin(\pi\alpha^k)}{x_0 + (1-x_0)\zeta^{\alpha^k} \cos(\pi\alpha^k)}\right) \right], \end{aligned} \quad (29)$$

where ρ and Π are given in Equation 15 and Equation A7, respectively, and $\zeta = \varepsilon/(1 - \varepsilon)$. If the initial reference class is

assumed to be the mean class, this is the probability that the mutation will be fixed or extinct in the mean class $k\langle\tau_k\rangle$ generations later. It will then require $q\langle\tau_k\rangle$ more generations before the mutation is fixed or extinct population-wide, corresponding to the time for the mean fitness class to go extinct. Given a mutation at a measured frequency x_0 at $t = 0$, the true, population-wide sojourn time in generations is then $\mathcal{P}_{\text{soj}}((k + q)\langle\tau_k\rangle|x_0) = \mathcal{P}_{\text{soj}}(k|x_0)$. Both simulated and predicted sojourn times for a number of different parameters are shown in Figure 6.

A similar method is used to calculate the expected time for fixation of any lineage in a given reference fitness class, which we dub $P_{\text{fix}}(k)$. Specifically, $P_{\text{fix}}(k)$ is the probability that one of the mutations founded in class 0 is past the threshold $1 - \varepsilon$, k fitness classes forward. Using the distribution of new mutations given in Equation 17, the probability that one is at frequency greater than $(1 - \varepsilon)$ in class k is

$$P_{\text{fix}}(k) = \frac{\sin(\pi\alpha^k)}{\pi} \int_{1-\varepsilon}^1 \frac{(1-x_k)^{\alpha^k-1}}{x_k^{\alpha^k+1}} dx_k \approx \varepsilon^{\alpha^k} \quad (30)$$

for $\varepsilon \ll 1$ and k large. In this case, the expected fixation time is

$$\langle k_{\text{fix}} \rangle = \sum_{k=1}^{\infty} k \frac{d\varepsilon^{\alpha^k}}{d\varepsilon} \approx q(\gamma_E + \log(\log(1/\varepsilon))) \sim \mathcal{O}(2q) \quad (31)$$

for q large, where the last relation holds for $\varepsilon \sim 10^{-2}$. Thus, it takes a time of about $2q\langle\tau_k\rangle$ for a mutation destined to fix to sweep in any class and an additional time of about $2q\langle\tau_k\rangle$ for that class to sweep through the population. The expected sweep time for successful mutations is then $\mathcal{O}(4q\langle\tau_k\rangle) \sim \mathcal{O}(4 \log(s/U_b)/s)$ for q large.

We note that the time for any mutation to cross some threshold frequency $1 - \varepsilon$ is closely related to the time it takes a sample of $n \gg 1$ individuals to coalesce within some fitness class, with $\varepsilon \propto 1/n$. In this case, for the coalescence time we recover the formula $q(\log(\log(n)) + \mathcal{O}(1))$ derived by both Desai *et al.* (2013) and Neher and Hallatschek (2013) for the corresponding backward process (for the latter work, the factor of q must be replaced by the model independent T_c , the coalescence time of two individuals in the high-fitness nose).

Discussion

We have used a simple, infinite-sites model of adaptation featuring a single beneficial selection coefficient to carefully account for the effects of genetic drift, mutation, selection, and by extension, genetic draft in determining the evolutionary dynamics of polymorphic sites. In what follows we discuss the connection between the dynamics of rapid adaptation and the Bolthausen–Sznitman coalescent, and comment on applications of our results for interpreting frequency trajectories and

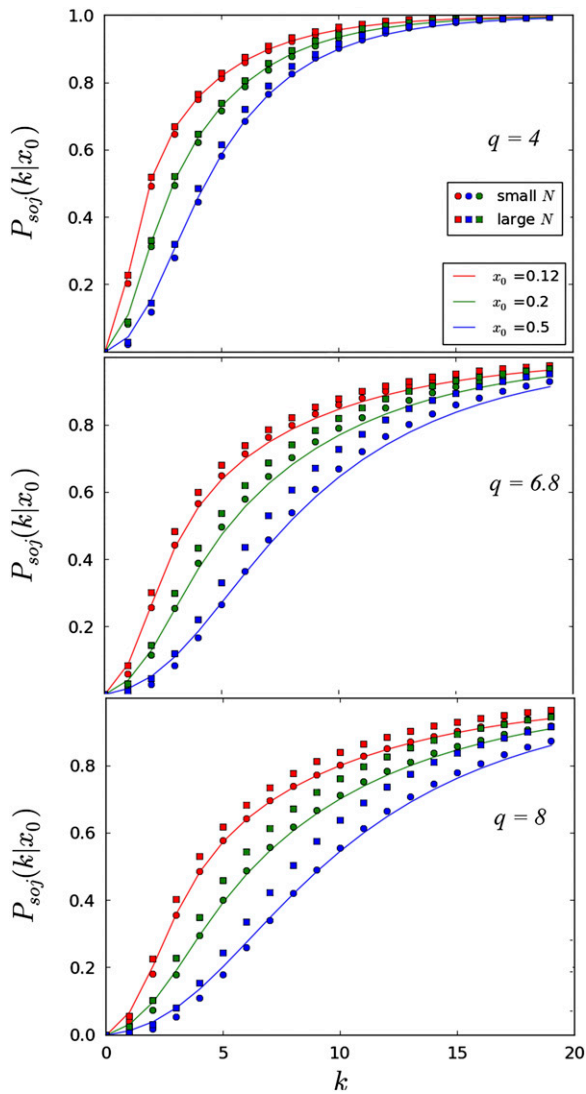


Figure 6 The sojourn probability $P_{soj}(k|x_0)$ plotted for different values of x_0 , q , and other population parameters U_b , s , and N . Curves represent the theoretical prediction for the sojourn times given in Equation 29. Circles and squares represent the results of forward-time simulations. The circular and square points have identical values of q , but differ in population size by a factor of 10. Circles represent simulations run with parameters $s = 0.1$, $U_b = 0.001$, and squares are run with parameters $s = 0.01$, $U_b = 0.0001$. The three plots represent different values of q , with $q = 4$ calculated from the approximate value in the text, and $q = 6.8, 8$ calculated with higher-order corrections given in Desai and Fisher (2007). For both simulated and theoretical curves, mutations were considered extinct or fixed in a given class when they fell below or rose above a cutoff of $\varepsilon = 0.03$ or $(1 - \varepsilon) = 0.97$, respectively. More information about the simulations may be found in File S1.

inferring the action of positive selection on sequence data. We also discuss pertinent similarities between our model and other models of genetic draft. Finally, we note that our method might be applied toward understanding evolution on populations with strong epistasis between evolutionary trajectories.

Rapid adaptation and the Bolthausen–Sznitman coalescent

Having established the effect of genetic draft on the frequencies of polymorphic sites, we then explored the effect that the stochastic jumps characterizing the process have on genetic diversity. In particular, we derived the site frequency spectrum of both beneficial and neutral mutations. As expected, we found that nearly private variants—which overwhelmingly occur and drift near the mean class—decay according to the well-known $1/x$ behavior predicted for unlinked neutral sites, whereas common mutations exhibit site frequency spectra more closely related to those of exponentially expanding clones, with the additional feature of an upswing at high frequency that is characteristic of adaptation for many models (Messer and Petrov 2013; Neher and Hallatschek 2013).

Of particular interest is the diversity of populations in the limiting case of $q \rightarrow \infty$, whose genealogies obey the Bolthausen–Sznitman coalescent. Note that in this case, the Bolthausen–Sznitman coalescence rates apply for individuals in the lead, with each “generation” taken over the time of a new fitness class establishment (elaboration on this setup can be found in Desai *et al.* 2013). Although this genealogical structure was observed previously in Desai *et al.* (2013), in the context of these stochastic jumps its origin is intuitive: in the high- q limit, the contribution of one individual at the wavefront to the next class falls off as $1/x^2$, which can be seen by taking $x_{k-1} \rightarrow 0$ in Equation 16. As proven by Schweinsberg (2003), the genealogies of populations with offspring distributions that decay as x^{-2} converge to the Bolthausen–Sznitman coalescent in the limit of $N \rightarrow \infty$. The $1/x^2$ decay is also well known to describe the frequency spectrum of exponentially expanding populations and in the case of adapting populations results from the exponential expansion of new mutants at the front of the wave. The Bolthausen–Sznitman coalescent has been associated with a growing number of different adaptive models (Brunet and Derrida 2012, 2013; Desai *et al.* 2013; Neher and Hallatschek 2013), suggesting that such genealogies may be a universal limiting feature of rapid adaptation.

Quasi-neutrality over long timescales

When many beneficial mutations segregate simultaneously, frequency dynamics of mutations begin to exhibit a qualitatively different behavior than those in the strong selection, weak mutation regime. After the introduction and establishment of a mutation in the fittest class, a mutation’s fixation probability is equal to its frequency in that class, regardless of its intrinsic fitness effect. To see this, assume a mutation first freezes to some frequency x_0 in its founding class. Then, since $x_{k \rightarrow \infty} \in \{0, 1\}$ (*i.e.*, at long times the allele is either fixed or extinct), this implies that the fixation probability

$$\begin{aligned}
P_{\text{fix}} &= \int_{1-1/N}^1 \rho(x_\infty | x_0) dx_\infty \\
&= 1 \left(\int_{1-1/N}^1 \rho(x_\infty | x_0) dx_\infty \right) + 0 \left(\int_0^{1-1/N} \rho(x_\infty | x_0) dx_\infty \right) \\
&= \langle x_\infty \rangle = x_0.
\end{aligned} \tag{32}$$

Essentially, this means that once a mutation founded at the high-fitness wavefront freezes to a particular frequency in a class, its likelihood of success depends only on its frequency in that class (at least, without more information about its frequency in fitter classes). Note, however, that the above formula technically holds only for mutations at high enough frequencies to have necessarily been founded near the distribution's nose class, since mutations founded away from the wavefront have fixation probabilities that are virtually zero. However, since these latecomers never reach more than a negligible frequency in their class, the above formula still describes the fixation probability of a randomly selected mutant reasonably well.

Now, once a mutation freezes in the fittest class, it will typically be present in the mean fitness class a time $q\langle\tau_k\rangle$ later. This corresponds to the timescale for the fittest class to become the mean class. Since the mean class consists of the majority of individuals in the population, the measured frequency of the mutant at this time is a good approximation to its frequency in the mean class. Barring any information about the frequency of the mutation in fitter classes, its measured frequency at this time is then roughly its fixation probability. (A better approximation is straightforwardly obtained by taking into account the fact that the mutation's frequency in classes below the mean is identically 0 at this time.) This equality becomes (nearly) exact at a time of $\mathcal{O}(2q\langle\tau_k\rangle)$, when the founding fitness class becomes the least fit in the population, and the mutation's frequency in all classes is determined solely by draft.

These considerations have important consequences for predicting the fate of polymorphic sites in experimentally evolving populations. In a population that is adapting rapidly enough for a stratification of fitnesses to always be present, sites that are polymorphic for sufficiently long have dynamics that are indistinguishable from neutral mutations at comparable frequencies in these populations, because they are determined *only* by genetic draft. Once the frequency of a mutation is frozen in a particular fitness class, its fitness effect alone is no longer important in determining its future success. This is because it is only the net fitness of the mutant *and* its genetic background that is important in determining its dynamics, and not the fitness effect of the mutation itself. Because it takes a time of $\mathcal{O}(2q\langle\tau_k\rangle)$ for a mutation founded at the nose to occupy all strata of fitnesses in the population, common mutations in these rapidly adapting populations can be thought to have a *relaxation time* of $2q\langle\tau_k\rangle$, beyond which their population-wide frequency approaches their frequency in the mean class, and the fitness of their founding haplotype decouples from their population-wide dynamics.

In short, the significance of this finding is simple: if a mutation—regardless of its fitness effect—has been mea-

surable in the population for longer than $2q\langle\tau_k\rangle$ generations, its fixation probability is equal to its frequency in the population. This approximation should already be quite good (albeit consistently too low due to the inclusion of less-fit classes where the mutation is absent) at a time $q\langle\tau_k\rangle$.

These considerations provide a parallel to Haldane's formula for the fixation probability of a beneficial mutation in the successive sweep regime, $P_{\text{fix}} = s$. Whereas this formula gives the likelihood of a mutation to fix despite the stochastic effects of genetic drift, the probability of a long-standing polymorphism to fix despite the randomizing effect of draft is simply $P_{\text{fix}} = x$, where x is the frequency of the mutation in the population. The two forces are similar in that they stochastically amplify the fluctuations in the trajectories of polymorphic sites, but the length- and timescale of the fluctuations caused by draft are much greater.

Implications for the McDonald–Kreitman test

Another useful application of our findings is the ability to analytically correct for the effect of genetic draft on the results of tests for signals of adaptation, such as the McDonald–Kreitman test (McDonald and Kreitman 1991). This test, along with many other widely used tests for selection, assumes that beneficial mutations are rare and segregate independently. However, both assumptions are invalid for rapidly adapting populations, and new analytical predictions are needed. Fortunately, we demonstrate that our method provides a simple way to correct for the effect of linkage between beneficial mutations. For the case of the McDonald–Kreitman test, this correction is straightforwardly obtained by accounting for the extra heterozygosity contributed by many simultaneously segregating beneficial mutations.

The McDonald–Kreitman test approximates the fraction α_{MK} of nucleotide substitutions that are adaptive by considering relative quantities of fixed and polymorphic, synonymous and nonsynonymous sites between two diverged populations. The fraction of adaptive substitutions is simply

$$\alpha_{\text{MK}} = \frac{d_+}{d_n} \approx 1 - \frac{d_s p_n}{d_n p_s}, \tag{33}$$

where d_+ , d_s , and d_n are the adaptive, synonymous, and nonsynonymous substitution rates, and p_n , p_s are the numbers of nonsynonymous and synonymous polymorphisms in one of the sampled populations, respectively. The last approximation arises from the assumption that

$$d_+ = d_n - \bar{d} \approx d_n - d_s(\bar{p}/p_s) \approx d_n - d_s(p_n/p_s), \tag{34}$$

where \bar{d} is the substitution rate of nonadaptive, nonsynonymous mutations and \bar{p} is the number of nonadaptive, nonsynonymous polymorphic sites in the sample. Implicit in Equation 34 are several assumptions: first, the rate of nonadaptive nonsynonymous substitutions in the sample is equal to the rate of synonymous substitutions, scaled by the relative frequencies of nonadaptive nonsynonymous to synonymous polymorphisms. This implicitly assumes that deleterious

mutations do not fix, that deleterious mutations do not significantly contribute to the measured numbers of nonsynonymous polymorphisms in the sample, and that the population has not undergone any demographic change to skew the distributions of polymorphic sites. Second, it is assumed that the number of nonadaptive, nonsynonymous polymorphisms is precisely equal to the measured numbers of nonsynonymous polymorphisms. In other words, beneficial mutations are rare and fix quickly upon arising; thus, they are rarely present in the population as polymorphisms.

Of course, these assumptions break down when deleterious or beneficial mutations significantly contribute to the number of polymorphic sites and when linkage between sites skews the relative frequencies of mutations. These issues usually result in a measured α_{MK} that severely underestimates the true fraction of adaptive substitutions. A large body of work has been put forward in an effort to correct for this skew (Andolfatto 2008; Charlesworth and Eyre-Walker 2008; Eyre-Walker and Keightley 2009; Messer and Petrov 2013). For example, introducing a low-frequency cutoff for measured polymorphisms significantly improves estimates of α_{MK} for the case of many weakly deleterious mutations (Fay *et al.* 2001; Charlesworth and Eyre-Walker 2008), since in the absence of genetic linkage few deleterious mutations will ever reach high frequencies. However, few studies have carefully analyzed the effect of linkage on α_{MK} , particularly the effect of linked beneficial mutations. One notable exception is the work of Messer and Petrov (2013), who used a sophisticated extension of the McDonald–Kreitman test accounting for demographic history and distributions of fitness effects to infer α_{MK} from simulated rapidly adapting populations. The authors found that the inference of a massive population expansion (derived from the site frequency spectrum of the sample) resulted in superior estimates of α_{MK} , although no such expansion ever occurred in the simulation. The intuition guiding this finding is that the site frequency spectrum of a population undergoing rapid adaptation resembles that of a population undergoing an exponential expansion. Regardless, analytic corrections for the effect of genetic draft have yet to be derived and would provide for a more straightforward way of accounting for this confounding factor.

Our results provide a simple analytical correction to α_{MK} for the case of tightly linked sections of the genome that accounts for genetic draft. First, we note that the number of polymorphic sites is closely related to heterozygosity π , the average number of nucleotide differences between two randomly drawn individuals, through

$$\pi = \left\langle \sum_{i=1}^p 2x_i(1-x_i) \right\rangle = 2p(\langle x \rangle - \langle x^2 \rangle). \quad (35)$$

Thus, the ratio p_n/p_s is simply

$$\frac{p_n}{p_s} = \frac{\pi_n(\langle x \rangle_s - \langle x^2 \rangle_s)}{\pi_s(\langle x \rangle_n - \langle x^2 \rangle_n)} \approx \frac{\pi_n}{\pi_s}. \quad (36)$$

In general, the moments of the frequencies of synonymous and nonsynonymous sites may be measured straightforwardly from the site frequency spectrum of the sample. However, as we have already demonstrated, in the rapid adaptation limit the SFS of neutral and beneficial mutations is similar in form. Thus, the heterozygosities of beneficial and neutral mutations are largely determined by different numbers of neutral and selected sites—*i.e.*, by p_n and p_s —rather than significantly different frequency distributions.

The heterozygosity for beneficial mutations is calculated from the moments of $\rho(x_k|x_0)$ (using the same method used in the calculation of the beneficial SFS) to be $\pi \approx 2(q-1)$, with higher-order corrections given in Desai *et al.* (2013). The neutral heterozygosity is simply $2U_nT_2$, where T_2 is the expected coalescence time for two randomly chosen individuals. The rate of beneficial substitutions is $1/\langle \tau_k \rangle$, and the rate of neutral substitutions is U_n . Thus, given the rate of synonymous mutations, $U_{n,s}$, and the rate of neutral nonsynonymous mutations, $U_{n,n}$, the expected, measured value of α_{MK} (by naively plugging in each measured value in Equation 33) will be

$$\begin{aligned} \langle \alpha_{MK,meas} \rangle &= 1 - \frac{d_s p_n}{d_n p_s} \approx 1 - \frac{d_s \pi_n}{d_n \pi_s} \\ &= 1 - \left(\frac{U_{n,s}}{1/\langle \tau_k \rangle + U_{n,n}} \right) \left(\frac{2T_2 U_{n,n} + 2(q-1)}{2T_2 U_{n,s}} \right) \end{aligned} \quad (37)$$

Clearly, only the first term of π_n corresponds to nonadaptive sites. Thus, we have

$$\begin{aligned} \langle \alpha_{MK,meas} \rangle &= \frac{d_+}{d_n} - \left(\frac{U_{n,s}}{1/\langle \tau_k \rangle + U_{n,n}} \right) \left(\frac{2(q-1)}{2T_2 U_{n,s}} \right) \\ &= \frac{d_+}{d_n} - \frac{2(q-1)d_s}{d_n \pi_s}. \end{aligned} \quad (38)$$

This gives for the fraction of adaptive substitutions

$$\frac{d_+}{d_n} = 1 - \frac{d_s p_n}{d_n p_s} + \frac{d_s 2(q-1)}{d_n \pi_s} = \langle \alpha_{MK,meas} \rangle + \frac{d_s 2(q-1)}{d_n \pi_s}. \quad (39)$$

In practice, the parameter q is measurable from estimates of N , s , and U_b or the distribution of fitnesses within the population.

The interpretation of this correction is intuitively simple. In populations where adaptation is rapid, the assumption that beneficial mutations do not contribute significantly to measured polymorphism breaks down. As a result, ascribing all measured nonsynonymous polymorphism p_n to neutral (or deleterious) mutations results in an underestimate of d_+/d_n . In fact, in the limit of infinitely rapid adaptation, $q \rightarrow \infty$, $\alpha_{MK,meas} \rightarrow 1/2$, in contrast to the *true* fraction of adaptive substitutions, $d_+/d_n \rightarrow 1$. Our model, then, provides a simple correction for this underestimate by predicting the expected fraction of observed nonsynonymous polymorphism

(relative to synonymous polymorphism) that arises from beneficial mutations.

Relation to other models of genetic draft

There is a natural connection between our model and the classic model of genetic draft between a neutral locus and a strongly selected locus, first studied by Gillespie (2000, 2001). In these seminal works, the diffusive random walk of a neutral allele is coupled with the stochastic process of a hitchhiking event occurring at rate R , which drives the neutral allele to either fixation or extinction. In Gillespie's model, the time required to fix the strongly selected allele relative to the time between substitutions is small enough that fixation/extinction is assumed to occur instantaneously. One basic result derived under the assumptions of such a model are the first two moments of the stochastic jump process: $\langle \Delta x \rangle = 0$, $\langle \Delta x^2 \rangle = Rx_0(1 - x_0)$, for $\Delta x = x_1 - x_0$ [here, x_1 denotes the frequency of the neutral allele after some time of $\mathcal{O}(\langle \tau_{\text{fix}} \rangle)$ for the strongly selected locus]. It turns out that the stochastic process described by Gillespie emerges naturally as the weak mutation limit of the model studied in this work.

To demonstrate this, we observe that the first two moments of our jump distribution $\rho(x_1|x_0)$ are $\langle \Delta x \rangle = 0$ and $\langle \Delta x^2 \rangle = x_0(1 - x_0)/q$ (the derivation, along with a general formula for higher order moments, is given in File S1). The similar functional dependence on x_0 between the two models is not particularly surprising, since it arises from a general property of any exchangeable coalescent (because the jump process describes the changing composition of each new lead class, and because all individuals within the lead have the same fitness, our model is exchangeable). However, the prefactor for each moment depends strongly on the stochastic dynamics predicted by different models. To compare the stochastic dynamics described by the different works, we first note that all of our results take time in units of fitness-class establishments, which is equal to the substitution time of beneficial mutations. In the weak mutation limit, we have $q \rightarrow 1$, so that $\langle \Delta x^2 \rangle \rightarrow x_0(1 - x_0) = Rx_0(1 - x_0)$, where R is the substitution rate of beneficial mutations. Higher-order moments of the two distributions are also similarly related, so that Gillespie's results emerge from our model in the limit that $q \rightarrow 1$. Thus, our results generalize the stochastic process describing genetic draft to the regime where the time between beneficial mutations is no longer large.

Next, we might ask how generalizable our results are when compared with genetic draft arising from different sources. For example, Neher and Shraiman (2011) recently analyzed the effect of draft arising from infrequent recombination events instead of *de novo* mutations. They introduced an approach similar to ours to study how the frequency of an allele changes over the course of a "meta-generation" set by a typical turnover timescale. The resulting dynamics are qualitatively similar to many features of our model. Specifically, Neher and Shraiman (2011) found that the origination of high-fitness clonal bubbles determines the long-term frequency dynamics of an allele and showed that

these dynamics can be encapsulated in an effective offspring distribution for recombinant genotypes, which decays asymptotically as x^{-2} . They then linked this effective offspring distribution to genealogies described by multiple merger coalescents. As a consequence, they found that the asymptotic form of the site frequency spectrum also decays asymptotically as x^{-2} for both neutral and beneficial alleles, analogous to our results here. The parallels between these two models reflect some general effects of draft on the dynamics of linked sites, which are robust to the specific source of the randomizing effect.

Applications to forks and more complicated fitness landscapes

Finally, because of the flexibility of our model with respect to the underlying fitness landscape, it is straightforward to extend our results to a small class of more general models involving complex interactions between sites. Specifically, we might consider the case in which two or more disjoint and incompatible evolutionary pathways are possible. That is, each individual can acquire one of two or more distinct sets of beneficial mutations, but mutations from different sets are incompatible. Note here that different evolutionary pathways are effectively incompatible if there is sign epistasis between them that is at least as strong as $\sim s$, so that any individual that acquires mutations from two or more pathways must leave the lead of the wave and is unlikely to fix as a result. In this scenario, we can apply our approach to analyzing the jumps from class to class for each different pathway separately. Different relative mutation rates for each pathway simply contribute a different prefactor in the exponent of the generating function for each contributing lineage (*i.e.*, replacing the $x_{\ell,k-1}$ with some relative rate μ_ℓ in Equation 12). This allows us to apply our method to answer questions about the probability that the population will follow any given path, its time to fixation, and the degree to which the population will simultaneously explore the other pathways.

These considerations provide a different perspective on the fates of populations evolving on rugged fitness landscapes, and particularly on the effect of a larger population size in avoiding local fitness peaks. Previous works have suggested that over certain timescales, smaller populations may have an advantage in adapting on these rugged landscapes, because their trajectories are more heterogeneous, whereas larger populations have an increased tendency to get stuck on local fitness peaks (Rozen *et al.* 2008; Handel and Rozen 2009; Jain *et al.* 2010; Szendro *et al.* 2013). However, our analysis suggests that if the landscape is dominated by several distinct uphill trajectories featuring mutational steps of similar size, large populations may be capable of traveling for many steps down multiple paths, effectively exploring the surrounding landscape before settling upon one particular uphill trajectory. For example, in the case of a simple fork with equal mutation rates down each pathway, a large, rapidly adapting population will typically explore about $2q$ mutational steps

forward before a particular pathway is closed off. The transition between this behavior and that considered in the work cited above evidently occurs between classical clonal interference, in which adaptation is dominated by the rare emergence of extremely fit mutants, and the multiple mutations regime, in which most fixed mutations are of roughly the same size. This, in turn, strongly depends on the distribution of fitness effects of mutations, with long-tailed or short-tailed distributions giving rise to dynamics dominated by clonal interference or multiple mutations, respectively (Desai and Fisher 2007; Fogle *et al.* 2008). The different outcomes predicted by these two regimes could explain the lack of experimental consensus on the effect of population size on outcomes of adaptation (Rozen *et al.* 2008; Schoustra *et al.* 2009; Miller *et al.* 2011).

Conclusions and Future Work

By using a simple model, we have made considerable headway in understanding how genetic draft affects the frequencies of mutations through a series of stochastic jumps, how these jumps affect genetic diversity, sojourn times, and fixation times of mutations, and why these statistics resemble those derived from the Bolthausen–Sznitman coalescent. We then showed how our method leads to a simple correction to the McDonald–Kreitman test that accounts for linkage between beneficial mutations. Finally, we discussed how our analysis might be extended to describe evolution on certain classes of rugged fitness landscapes, which—although admittedly very simple—nonetheless describe limiting behavior for sign epistasis between multiple evolutionary pathways.

Still, our model has some shortcomings. First, we neglect recombination, making our results applicable only to the evolution of microbial populations and tightly linked regions of the genomes of sexually reproducing organisms. Naturally, in cases where recombination is no longer rare, the effects of genetic draft are tempered as competing beneficial mutations recombine onto a single genetic background. Fitness classes that evolve disjointly in the asexual model are then allowed to mingle at each reproductive step, meaning that a series of stochastic jumps between classes no longer correctly describe the dynamics. Fortunately, the work of Neher and Shraiman (2011), which accounts for the effects of occasional (facultative) outcrossing of clones provides a framework for combining these two sources of genetic draft. In particular, since common mutations must at one point propagate near the most-fit class, evolutionary dynamics in these populations are still largely informed by the distribution of haplotypes in the nose. This distribution would then obtain contributions from both mutations from the adjacent class and recombined haplotypes obtained from mating between less-fit clones.

We also make use of the assumption of a single selection coefficient. Indeed, two facets of our model that are key in deriving analytical results—the organization of clones according to fitness classes and the asymptotic freezing of

frequencies in each class—both break down when a single selection coefficient is replaced with some distribution of fitness effects. However, several works (Desai and Fisher 2007; Good *et al.* 2012) have shown that even in populations with a distribution of fitness effects, evolutionary dynamics are well described by the use of an effective, or *predominant* selection coefficient, which coincides exactly with the most common fixed mutational effect. Still, the inclusion of a distribution of fitness effects, and the resulting unified understanding of the effects of multiple mutations and mutations of varying effect sizes in driving evolutionary dynamics, remains a promising subject of future work.

Acknowledgments

We thank Benjamin Good, Sergey Kryazhimskiy, Richard Neher, and an anonymous reviewer for their helpful comments and suggestions. This work was supported by a National Science Foundation Graduate Research Fellowship (K.K.) and by the James S. McDonnell Foundation, the Alfred P. Sloan Foundation, and grant GM104239 from the National Institutes of Health (M.M.D.).

Literature Cited

- Andolfatto, P., 2008 Controlling type-I error of the McDonald–Kreitman test in genomewide scans for selection on noncoding DNA. *Genetics* 180: 1767–1771.
- Brunet, E., and B. Derrida, 2012 How genealogies are affected by the speed of evolution. *Philos. Mag.* 92: 255–271.
- Brunet, E., and B. Derrida, 2013 Genealogies in simple models of evolution. *J. Stat. Phys.* P01006.
- Brunet, E., I. M. Rouzine, and C. O. Wilke, 2008 The stochastic edge in adaptive evolution. *Genetics* 179: 603–620.
- Charlesworth, J., and A. Eyre-Walker, 2008 The McDonald–Kreitman test and slightly deleterious mutations. *Mol. Biol. Evol.* 25: 1007–1015.
- de Visser, J. A., and D. E. Rozen, 2006 Clonal interference and the periodic selection of new beneficial mutations in *Escherichia coli*. *Genetics* 172: 2093–2100.
- de Visser, J. A., C. W. Zeyl, P. J. Gerrish, J. L. Blanchard, and R. E. Lenski, 1999 Diminishing returns from mutation supply rate in asexual populations. *Science* 283: 404–406.
- Der, R., C. Epstein, and J. B. Plotkin, 2012 Dynamics of neutral and selected alleles when the offspring distribution is skewed. *Genetics* 191: 1331–1344.
- Desai, M., and D. Fisher, 2007 Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics* 176: 1759.
- Desai, M., A. Walczak, and D. Fisher, 2013 Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics* 193: 565–585.
- Eldon, B., and J. Wakeley, 2006 Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* 172: 2621–2633.
- Eyre-Walker, A., and P. D. Keightley, 2009 Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* 26: 2097–2108.
- Fay, J. C., and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405–1413.

- Fay, J. C., G. J. Wyckoff, and C.-I. Wu, 2001 Positive and negative selection on the human genome. *Genetics* 158: 1227–1234.
- Fisher, D. S., 2013 Asexual evolution waves: fluctuations and universality. *J. Stat. Mech.* P01011.
- Fogle, C. A., J. L. Nagle, and M. M. Desai, 2008 Clonal interference, multiple mutations and adaptation in large asexual populations. *Genetics* 180: 2163–2173.
- Gerrish, P. J., and R. E. Lenski, 1998 The fate of competing beneficial mutations in an asexual population. *Genetica* 102–103: 127–144.
- Gillespie, J. H., 2000 Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* 155: 909–919.
- Gillespie, J. H., 2001 Is the population size of a species relevant to its evolution? *Evolution* 55: 2161–2169.
- Good, B. H., I. M. Rouzine, D. J. Balick, O. Hallatschek, and M. M. Desai, 2012 Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proc. Natl. Acad. Sci. USA* 109: 4950–4955.
- Hallatschek, O., 2011 The noisy edge of traveling waves. *Proc. Natl. Acad. Sci. USA* 108: 1783–1787.
- Handel, A., and D. E. Rozen, 2009 The impact of population size on the evolution of asexual microbes on smooth vs. rugged fitness landscapes. *BMC Evol. Biol.* 9: 236.
- Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton *et al.*, 2011 Classic selective sweeps were rare in recent human evolution. *Science* 331: 920–924.
- Jain, K., J. Krug, and S.-C. Park, 2010 Evolutionary advantage of small populations on complex fitness landscapes. *Evolution* 65: 1945–1955.
- Kao, K. C., and G. Sherlock, 2008 Molecular characterization of clonal interference during adaptive evolution in asexual populations of *Saccharomyces cerevisiae*. *Nat. Genet.* 40: 1499–1504.
- Karasov, T., P. W. Messer, and D. A. Petrov, 2010 Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. *PLoS Genet.* 6: e1000924.
- Kim, Y., and H. A. Orr, 2005 Adaptation in sexuals vs. asexuals: clonal interference and the Fisher–Muller model. *Genetics* 171: 1377–1386.
- Kim, Y., and W. Stephan, 2003 Selective sweeps in the presence of interference among partially linked loci. *Genetics* 164: 389–398.
- Lang, G. I., D. Botstein, and M. M. Desai, 2011 Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics* 188: 647–661.
- Lang, G. I., D. P. Rice, M. J. Hickman, E. Sodergren, G. M. Weinstock *et al.*, 2013 Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* 500: 571–574.
- McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the ADH locus in *Drosophila*. *Nature* 351: 652–654.
- McVean, G. A. T., and B. Charlesworth, 2000 The effects of Hill–Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* 155: 929–944.
- Messer, P. W., and D. A. Petrov, 2013 Frequent adaptation and the McDonald–Kreitman test. *Proc. Natl. Acad. Sci. USA* 110: 8615–8620.
- Miller, C. R., P. Joyce, and H. A. Wichman, 2011 Mutational effects and population dynamics during viral adaptation challenge current models. *Genetics* 187: 185–202.
- Miralles, R., P. J. Gerrish, A. Moya, and S. F. Elena, 1999 Clonal interference and the evolution of RNA viruses. *Science* 285: 1745–1747.
- Muller, H., 1932 Some genetic aspects of sex. *Am. Nat.* 66: 118–138.
- Neher, R. A., and O. Hallatschek, 2013 Genealogies of rapidly adapting populations. *Proc. Natl. Acad. Sci. USA* 110: 437–442.
- Neher, R., and B. Shraiman, 2011 Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics* 188: 975–996.
- Neher, R. A., B. I. Shraiman, and D. S. Fisher, 2010 Rate of adaptation in large sexual populations. *Genetics* 184: 467–481.
- Park, S.-C., and J. Krug, 2007 Clonal interference in large populations. *Proc. Natl. Acad. Sci. USA* 104: 18135–18140.
- Park, S.-C., D. Simon, and J. Krug, 2010 The speed of evolution in large asexual populations. *J. Stat. Phys.* 138: 381.
- Pritchard, J. K., J. K. Pickrell, and G. Coop, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* 20: R208–R215.
- Rouzine, I., J. Wakeley, and J. M. Coffin, 2003 The solitary wave of asexual evolution. *Proc. Natl. Acad. Sci. USA* 100: 587–592.
- Rouzine, I. M., E. Brunet, and C. O. Wilke, 2008 The traveling wave approach to asexual evolution: Muller’s ratchet and speed of adaptation. *Theor. Popul. Biol.* 73: 24–46.
- Rozen, D. E., M. G. Habets, A. Handel, and J. A. G. M. de Visser, 2008 Heterogeneous adaptive trajectories of small populations on complex fitness landscapes. *PLOS ONE* 3:e1715.
- Schoustra, S. E., T. Bataillon, D. R. Gifford, and R. Kassen, 2009 The properties of adaptive walks in evolving populations of fungus. *PLoS Biol* 7: e1000250.
- Schweinsberg, J., 2003 Coalescent processes obtained from supercritical Galton–Watson processes. *Stochast. Process. Appl* 160: 107–139.
- Sella, G., D. A. Petrov, M. Przeworski, and P. Andolfatto, 2009 Pervasive natural selection in the *Drosophila* genome? *PLoS Genet* 5: e1000495.
- Smith, J. M., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. *Genet. Res.* 23: 23–35.
- Szendro, I. G., J. Franke, J. A. G. M. de Visser, and J. Krug, 2013 Predictability of evolution depends nonmonotonically on population size. *Proc. Natl. Acad. Sci. USA* 110: 571–576.
- Tsimring, L., H. Levine, and D. Kessler, 1996 RNA virus evolution via a fitness-space model. *Phys. Rev. Lett.* 76: 4440–4443.
- Wright, S., 1938 The distribution of gene frequencies under irreversible mutation. *Proc. Natl. Acad. Sci. USA* 24: 253–259.

Communicating editor: L. M. Wahl

Appendix: Dynamics of the Transition Process

In this appendix we show that frequencies of mutant lineages are frozen when a class begins feeding mutants to the lead that are destined to establish. We then explicitly derive the probability distribution ρ_k of a transition in a mutation's frequency from a starting class 0 to some final fitness class k .

To prove that frequencies of mutant lineages are frozen when a class begins supplying establishing mutants to the next class, we first note that the L th establishing mutant in a given fitness class typically occurs at time t_L such that

$$L = \underbrace{qs}_{(a)} \cdot U_b \underbrace{\int_0^{t_L} \frac{e^{(q-1)st}}{qs} dt}_{(b)}, \quad (\text{A1})$$

where (a) is the establishment probability of one mutant and (b) is the total number of mutants introduced into the lead class by time t_L . Thus, using the same argument as in Desai *et al.* (2013), the amount that the L th establishing lineage contributes to a fitness class as a fraction of the first lineage is

$$\frac{\eta_{L,k}(t)}{\eta_{1,k}(t)} = \frac{e^{qs(t-t_L)}}{e^{qs(t-t_1)}} = \frac{1}{L^{q/(q-1)}}. \quad (\text{A2})$$

Note that this is an *upper* limit on the contribution of $\eta_{L,k}$, since the growth of each subsequent lineage actually decreases according to the rate of adaptation v . A fitness class typically establishes in a time $\langle \tau_k \rangle$ and generates its first establishing mutant a time t_1 after that. If we neglect the decreasing growth rate due to adaptation of the population (valid for large q), then at this point, the class below it has supplied

$$qsU_b \int_0^{t_1 + \langle \tau_k \rangle} \frac{e^{(q-1)st}}{qs} dt \sim \frac{s}{U_b} \gg 1 \quad (\text{A3})$$

establishing mutants.

In practice, however, few biological populations evolve with $q > 4$, in which case the diminishing growth rate becomes important in the above calculation. A simple modification of the analysis gives the number of establishing lineages to be roughly

$$\left((q-2)(q-1)^{(q-3)/(q-1)} / ((q-3)q) \right) (s/U_b)^{(q-2)/(q-1)}$$

for $q > 3$, which, although considerably smaller than the asymptotic value of s/U_b is offset by the more rapid decay of $\eta_{L,k}/\eta_{1,k}$ for smaller q . Extensions for $q = 2, 3$ are likewise straightforward.

Thus, we have demonstrated that the contribution of subsequent mutations diminishes rapidly, and by the time a fitness class begins feeding establishing mutants to the next class, it has $\mathcal{O}(s/U_b) \gg 1$ mutations that are destined to establish in it. As a result, the contribution of subsequent mutations after this time is already very small, meaning that the frequencies of common lineages in the fitness class at this time may safely be treated as frozen.

The astute reader might also note that mutant lineages that are destined to go extinct may also contribute to shifting the frequencies, which is certainly a contributing factor close to the time that the class establishes. Although not as straightforward a calculation, if $s \gg U_b$ then it is still true that by the time a class begins feeding establishing mutations, the contribution of these "doomed lineages" is also small (File S1). Thus we are justified in treating the process as a Markov chain.

Now we are ready to derive the transition probability of a mutation at frequency x_0 in class 0 to some frequency x_k in class k . Desai and Fisher (2007, Equation 29) showed that the contribution to class k , ν_k , from a lineage at frequency x_{k-1} in class $k-1$ is described by the generating function

$$\langle e^{-\nu_k z} \rangle = e^{-x_{k-1} z^\alpha}. \quad (\text{A4})$$

Thus, after class k establishes, the lineage grows as

$$\eta_k(t) = \frac{x_k}{qs} e^{(q-1)s(t-\tau_k)} \propto \frac{\nu_k}{qs} e^{(q-1)s(t-\langle \tau_k \rangle)}. \quad (\text{A5})$$

The generating function for the contribution ν_{k+1} of the mutation 2 fitness steps forward is then obtained by averaging over the intermediate value ν_k ,

$$\begin{aligned}
\langle e^{-\nu_{k+1}z} \rangle &= \langle e^{-\nu_k z^\alpha} \rangle_{\nu_k} = \int_0^\infty e^{-\nu_k z^\alpha} P(\nu_k | x_{k-1}) d\nu_k \\
&= \frac{1}{2\pi i} \int_0^\infty \int_{-i\infty}^{i\infty} e^{-\nu_k z^\alpha} e^{\nu_k \zeta} e^{-x_{k-1} \zeta^\alpha} d\zeta d\nu_k \\
&= \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{e^{-x_{k-1} \zeta^\alpha}}{z^\alpha - \zeta} d\zeta \\
&= e^{-x_{k-1} z^{\alpha^2}},
\end{aligned} \tag{A6}$$

where the step in line 2 makes use of the standard formula for inverting the Laplace transform. Although technically the average should be taken over the frequency x_k (which is analytically intractable), an average over ν_k is a reasonable approximation if σ_k , the growth of the entire class, does not deviate too much from its typical value, and if the growth of the lineage ν_k in future classes may be taken independently of the growth of other lineages. Because the effect of this approximation is compounded at each step forward, it introduces significant deviations at timescales of roughly q fitness steps forward, which is the timescale over which fluctuations in the advance of the fitness wave become significant. Accepting this approximation, it is then straightforward to show that

$$\langle e^{-\nu_k z} \rangle = e^{-x_0 z^{\alpha^k}}. \tag{A7}$$

Without loss of generality, we can index the class in which the mutant begins to be tracked to 0. The fitness class at $k = 1$ can then be divided into those individuals descended from a particular lineage in class 0, at frequency x_0 , and those not descended from that lineage, at frequency $(1 - x_0)$. Since, for each subsequent fitness class, new frequencies are frozen near the nose, where the two sets of individuals proliferate independently, there are two independent variables encoding the fate of the lineage: ν_k , denoting the contribution of the lineage to a class with k more beneficial mutations, and $\tilde{\nu}_k$, denoting the contribution of individuals not derived from that lineage, so that $\tilde{\nu}_k + \nu_k = \sigma_k$.

If we denote the probability densities of ν_k and $\tilde{x}_k = \tilde{\nu}_k / \nu_k$ by $P_k(\nu_k | x_0)$ and $\Pi_k(\tilde{x}_k | x_0)$, respectively, then

$$\begin{aligned}
\Pi_k(\tilde{x}_k | x_0) &= \int_0^\infty P_k(\tilde{x}_k \nu_k | 1 - x_0) P_k(\nu_k | x_0) \nu_k d\nu_k \\
&= \frac{1}{(2\pi i)^2} \int_0^\infty \int_{\varepsilon - i\infty}^{\varepsilon + i\infty} \int_{-i\infty}^{i\infty} e^{\tilde{x}_k \nu_k z_1} e^{\nu_k z_2} e^{-(1-x_0)z_1^{\alpha^k}} e^{-x_0 z_2^{\alpha^k}} \nu_k dz_1 dz_2 d\nu_k, \\
&= \frac{1}{(2\pi i)^2} \int_{\varepsilon - i\infty}^{\varepsilon + i\infty} \int_{-i\infty}^{i\infty} \frac{e^{-(1-x_0)z_1^{\alpha^k}} e^{-x_0 z_2^{\alpha^k}}}{(\tilde{x}_k z_1 + z_2)^2} dz_1 dz_2, \\
&= -\frac{x_0 \alpha^k}{2\pi i (-\tilde{x}_k)^{1-\alpha^k}} \int_{\varepsilon - i\infty}^{\varepsilon + i\infty} \frac{e^{-(1-x_0)z_1^{\alpha^k} - x_0 (-\tilde{x}_k z_1)^{\alpha^k}}}{z_1^{1-\alpha^k}} dz_1, \\
&= \frac{x_0 \alpha^k}{\pi} \operatorname{Re} \left[\int_0^\infty \frac{e^{-(1-x_0)(iz)^{\alpha^k} - x_0 (-\tilde{x}_k iz)^{\alpha^k}}}{(-i\tilde{x}_k z)^{1-\alpha^k}} dz \right], \\
&= -\frac{x_0}{\tilde{x}_k^{1-\alpha^k} \pi} \operatorname{Re} \left[\frac{1}{i(x_0 \tilde{x}_k^{\alpha^k} + (1-x_0)e^{i\pi \alpha^k})} \right],
\end{aligned} \tag{A8}$$

where $\varepsilon \in \mathbb{R}$, $\varepsilon \rightarrow 0^-$, and step 4 to 5 makes use of a transformation $z = -iz_1 + i\varepsilon$. Simplifying gives

$$\Pi_k(\tilde{x}_k | x_0) = \frac{\sin(\pi \alpha^k)}{\tilde{x}_k \pi} \frac{x_0(1-x_0)}{(1-x_0)^2 \tilde{x}_k^{-\alpha^k} + x_0^2 \tilde{x}_k^{\alpha^k} + 2x_0(1-x_0)\cos(\pi \alpha^k)}. \tag{A9}$$

Finally, we perform a change of variables to $x_k = 1/(1 + \tilde{x}_k)$, giving the jump distribution,

$$\rho_k(x_k | x_0) = \frac{\sin(\pi \alpha^k) x_0(1-x_0)}{x_k(1-x_k) \pi \left[(1-x_0)^2 \left(\frac{x_k}{1-x_k}\right)^{\alpha^k} + x_0^2 \left(\frac{1-x_k}{x_k}\right)^{\alpha^k} + 2x_0(1-x_0)\cos(\pi \alpha^k) \right]}. \tag{A10}$$

This function denotes the probability density of observing a mutation at frequency x_k in class k , given that it was at frequency x_0 in class 0. The function cited in the text is $\rho(x_1 | x_0) = \rho_1(x_1 | x_0)$.

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.156430/-/DC1>

The Dynamics of Genetic Draft in Rapidly Adapting Populations

Katya Kosheleva and Michael M. Desai

FILE S1

SUPPLEMENTARY INFORMATION

Contribution of doomed lineages to a fitness class:

In deriving the transition probabilities, we argue that the frequency of competing mutant lineages is frozen by the time a fitness class begins feeding mutants that are destined to establish into the nose. We arrive at this conclusion by calculating the number of establishing mutant lineages by the (typical) time this occurs. Since this number is typically large, and the contribution of each subsequent establishing lineage vanishes (despite the fact that these lineages arrive more and more quickly), the conclusion follows that the frequencies of lineages by this timepoint are frozen. This comes from the fact that the deterministic lineage size fraction, $n_L/n_1 \sim 1/L^{q/(q-1)}$ is convergent.

What could be the contribution of mutants that will not establish in a class, but may still contribute establishing mutants to the *next* fitness class? There are many more mutants introduced that will not establish relative to those that will, and these mutants grow roughly proportionally to the size of the class below them (i.e., at rate $(q-1)s$). By the time the population starts supplying establishing mutants, if these doomed lineages still comprise a significant fraction of the population, the result will be a deterministic drifting of the frequency, as these doomed lineages (which, because they are plentiful, will typically be split as x_0) become less and less of a contributing factor. We would like to determine if these combined lineages are negligible by the time a class begins feeding establishing mutants or still constituting some non-trivial fraction of the class.

The total number of individuals at time t derived from lineages that are destined to go extinct is given by

$$N_{doomed}(t) = \int_0^t \underbrace{dT U_b(1-qs) \frac{1}{qs} e^{(q-1)sT}}_{(a)} \underbrace{\langle n(t-T) | n(t \rightarrow \infty) = 0 \rangle}_{(b)}, \quad (1)$$

where (a) is the expected number of non-establishing mutants occurring in a small interval $d\tau$, and (b) is the expected number of these mutants that still persist at a time $t - \tau$ later, given that these mutants are going to eventually go extinct. Here, $\langle n(t - \tau) | n(t \rightarrow \infty) = 0 \rangle$ is derived from standard branching process analyses.

The probability distribution of doomed (but not yet extinct) lineages at time t is given by

$$P(n > 0, t|\text{doomed}) = P_{ext}(n)P(n, t)/P_{ext}(t), \quad (2)$$

where $P_{ext}(n)$ is the probability of extinction of a lineage composed of n individuals, $P_{ext}(t)$ is the probability that a lineage destined to go extinct is not yet extinct by time t , and $P(n, t)$ is the probability that a single mutant created at $t = 0$ has n descendants at time t . Note that by a “doomed” lineage we mean a lineage that is destined to vanish.

Now, following standard branching process analysis, a new mutant at fitness qs will have the following distribution of ancestors a time t later (Desai and Fisher, 2007; Eq. 11):

$$P(n = 0, t) = \frac{e^{qst} - 1}{(1 + qs)e^{qst} - 1}, \quad (3)$$

so that

$$P(n = 0, t|\text{doomed}) = \frac{(1 + qs)(e^{qst} - 1)}{(1 + qs)e^{qst} - 1}, \quad (4)$$

and furthermore (Desai and Fisher, 2007; Eq. 10)

$$P(n > 0, t) = \frac{(qs)^2 e^{qst}}{((1 + qs)e^{qst} - 1)((1 + qs)e^{qst} - 1 - qs)} \left(\frac{((1 + qs)e^{qst} - 1 - qs)}{(1 + qs)e^{qst} - 1} \right)^n. \quad (5)$$

In the branching process analysis, all lineages are independent, so

$$P_{ext}(n) = P_{ext}(1)^n = \left(\frac{1}{1 + qs} \right)^n. \quad (6)$$

Furthermore,

$$P_{ext}(t) = 1 - P(n = 0, t|\text{doomed}) = \frac{qs}{(1 + qs)e^{qst} - 1}. \quad (7)$$

Thus, using Eq. (2) gives

$$P(n > 0, t|\text{doomed}) = \frac{qse^{qst}}{(1 + qs)e^{qst} - 1 - qs} \left[\frac{e^{qst} - 1}{(1 + qs)e^{qst} - 1} \right]^n. \quad (8)$$

Finally, this gives us

$$\langle n(t) | n(t) \neq 0, n(t \rightarrow \infty) = 0 \rangle = \sum_{n=1}^{\infty} n P(n, t|\text{doomed}) = \frac{qse^{qst}}{(1 + qs)e^{qst} - 1 - qs} \sum_{n=1}^{\infty} n \xi^n \quad (9)$$

with

$$\xi = \frac{e^{qst} - 1}{(1 + qs)e^{qst} - 1}. \quad (10)$$

This reduces to

$$\langle n(t) | n(t) \neq 0, n(t \rightarrow \infty) \rangle = \frac{((1 + qs) - e^{-qst})}{qs} \quad (11)$$

Now, we include the possibility that the lineage has gone extinct by time t :

$$\langle n(t) | n(t \rightarrow \infty) = 0 \rangle = \frac{((1 + qs) - e^{-qst})}{qs} P[n \neq 0 | \text{doomed}] = e^{-qst}. \quad (12)$$

A fitness class, naively, has $\mathcal{O}(1/U_b)$ individuals when it typically starts creating establishing mutants. Hence,

$$\begin{aligned} \frac{N_{\text{doomed}}(t_1 + \tau_k)}{n_k((t_1 + \tau_k))} &= U_b^2 \int_0^{(t_1 + \tau_k)} (1 - qs) \frac{1}{qs} e^{(q-1)sT} e^{-qs((t_1 + \tau_k) - T)} dT \\ &\approx \frac{U_b^2 e^{qs(t_1 + \tau_k)}}{2qs} \\ &\approx \frac{s}{2}. \end{aligned} \quad (13)$$

Thus, given our assumption that $s \ll 1$, the contribution of doomed lineages is indeed small and can be ignored.

Moments of the jump probability: The moments of the jump probability ρ are calculated as follows, analogous to the method used in Desai et al. (2013). Using the fact that

$$\left(\frac{1}{\sigma_k}\right)^n = \int_0^\infty \frac{z^{n-1}}{(n-1)!} e^{-z\sigma_k} dz, \quad (14)$$

the n -th moment (for the step $k \rightarrow k + 1$) is derived as

$$\begin{aligned} \langle x^n \rangle &= \left\langle \left(\frac{\nu}{\sigma_k}\right)^n \right\rangle = \left\langle \int_0^\infty \frac{z^{n-1}}{(n-1)!} e^{-z\sigma_k} \nu^n dz \right\rangle \\ &= \int_0^\infty \frac{z^{n-1}}{(n-1)!} \langle e^{-z\sigma_k} \nu^n \rangle dz \\ &= \int_0^\infty \frac{z^{n-1}}{(n-1)!} \langle e^{-z(\nu + \bar{\nu})} \nu^n \rangle dz \end{aligned}$$

$$\begin{aligned}
&= \int_0^\infty \frac{z^{n-1}}{(n-1)!} \langle e^{-z\bar{\nu}} \rangle \langle e^{-z\nu} \nu^n \rangle dz \\
&= \int_0^\infty \frac{z^{n-1}}{(n-1)!} e^{-(1-x_0)z^{(1-1/q)}} \langle e^{-z\nu} \nu^n \rangle dz.
\end{aligned} \tag{15}$$

Now,

$$\langle \nu^n e^{-z\nu} \rangle = (-1)^n \frac{d^n}{dz^n} \langle e^{-z\nu} \rangle \tag{16}$$

and

$$\langle e^{-z\nu} \rangle = e^{-x_0 z^{(1-1/q)}}. \tag{17}$$

We can thus compute a general formula for the n -th derivative:

$$\frac{d^n}{dz^n} e^{-x_0 z^{(1-1/q)}} = e^{-x_0 z^{(1-1/q)}} \sum_{i=0}^{n-1} (-x_0)^{n-i} (1-1/q)^{n-i} z^{-(n-i)/q-i} (-1)^i \frac{\Gamma(1/q+i)}{\Gamma(1/q)}. \tag{18}$$

Thus,

$$\langle \nu^n e^{-z\nu} \rangle = e^{-x_0 z^{1-1/q}} \sum_{i=0}^{n-1} x_0^{n-i} (1-1/q)^{n-i} z^{-(n-i)/q-i} \frac{\Gamma(1/q+i)}{\Gamma(1/q)}. \tag{19}$$

This gives for the moments,

$$\begin{aligned}
\langle x^n \rangle &= \left\langle \left(\frac{\nu}{\sigma_k} \right)^n \right\rangle = \int_0^\infty \frac{z^{(n-1)}}{(n-1)!} e^{-(1-x_0)z^{1-1/q}} e^{-x_0 z^{1-1/q}} \sum_{i=0}^{n-1} x_0^{n-i} (1-1/q)^{n-i} z^{-(n-i)/q-i} \frac{\Gamma(1/q+i)}{\Gamma(1/q)} dz \\
&= \sum_{i=0}^{n-1} x_0^{n-i} (1-1/q)^{n-i} \frac{\Gamma(1/q+i)}{(n-1)! \Gamma(1/q)} \int_0^\infty z^{(n-1)} e^{-z^{1-1/q}} z^{-(n-i)/q-i} dz \\
&= \sum_{i=0}^{n-1} x_0^{n-i} \frac{\Gamma(1/q+i) \Gamma(n-i)}{(n-1)! \Gamma(1/q)} \left(\frac{q-1}{q} \right)^{n-1-i}.
\end{aligned} \tag{20}$$

The first two moments are

$$\langle \Delta x \rangle = 0, \tag{21}$$

$$\langle (\Delta x)^2 \rangle = \frac{x_0(1-x_0)}{q}. \tag{22}$$

Equation (20) is readily generalized to k steps forward through the usual substitution,

$1 - 1/q = \alpha \Rightarrow \alpha^k$, which gives

$$\langle x_k^n \rangle = \sum_{i=0}^{n-1} x_0^{n-i} \frac{\Gamma(1 - \alpha^k + i)\Gamma(n - i)}{(n - 1)!\Gamma(1 - \alpha^k)} (\alpha^k)^{n-1-i}. \quad (23)$$

Site Frequency Spectrum of Nearly Private Variants:

In this supplement we flesh out the derivation of the site frequency spectrum of nearly private variants. As noted in the text, several of the difficulties that arise in the derivation of the site frequency spectrum of common mutations may be neglected when dealing with rare variants. Specifically, the problem is vastly simplified with the inclusion of two approximations. First, since the lineage sizes of extremely rare variants are small and destined to go extinct, the lineages can be assumed to experience no further (establishing) beneficial mutations (specifically, this approximation holds if $nyU_b \ll 1$, where n is the lineage size and y the fitness of the mutant created by the lineage). Second, we assume that mutations are fed into a given fitness class k deterministically at rate $U_b n_{k-1}(t)$. This holds if $U_b n_{k-1}(t) \gg 1$ (certainly true in the bulk of the distribution), in which case fluctuations of incoming mutants around the expected number are small.

Because these lineages never comprise a significant fraction of the population, they can be studied through a standard branching process analysis with a constant death rate $d = 1$ and a time-varying birth rate $b(t) = 1 + y_0 - vt$, where t is time in generations, y_0 is the initial fitness of the mutant, and v the mean rate of adaptation of the population. First, we are interested in deriving the expected (time-averaged) number of mutations carried by n individuals with relative fitnesses y , $F_{rare}(n/N, y)$. This is obtained by considering the expected number of mutants introduced when a given fitness class was at relative fitness $y_0 > y$, and multiplying by the probability that, in the time it took for the relative fitness of the class to decrease to y , the lineage size of any of these mutants has increased to n . $F_{rare}(n/N, y)$ is then the integral over all possible landing fitnesses y_0 .

Clearly, the number of mutants introduced at an initial fitness y_0 is simply $U_b N(y_0 - s)$ where $N(y)$ is the expected number of individuals at relative fitness y . So long as the individual does not reside in the distribution's high-fitness nose, $N(y)$ is well approximated

by a Gaussian with variance v (Desai and Fisher, 2007):

$$N(y) = \frac{N}{\sqrt{2\pi v}} e^{-y^2/2v}. \quad (24)$$

Furthermore, a mutant that is introduced at initial fitness y_0 will be at fitness y in a time $t = (y_0 - y)/v$. Finally, the distribution in lineage sizes of a mutant with an initial birth rate $1 + y_0$ that is decreasing at a rate v per generation is a classic branching process problem that was solved by Kendall (1948). The distribution in lineage sizes is

$$P_1(n > 0, t) = \frac{e^{-y_0 t + \frac{v t^2}{2}}}{\left(\int_0^t e^{-y_0 \tau + v \tau^2 / 2} d\tau + 1\right)^2} \left(1 - \frac{e^{-y_0 t + \frac{v t^2}{2}}}{\int_0^t e^{-y_0 \tau + v \tau^2 / 2} d\tau + 1}\right)^{n-1}. \quad (25)$$

As a result, $F_{rare}(n/N, y)$ is given by

$$\begin{aligned} F_{rare}(n/N, y) &= \frac{1}{v} \int_y^{(q-2)s} P_1(n, (y_0 - y)/v) U_b N(y_0 - s) dx_0 \\ &= \frac{1}{v} \frac{U_b N}{\sqrt{2\pi v}} \int_y^\infty \frac{e^{\frac{y^2 + 2y_0 s - 2y_0^2 - s^2}{2v}}}{\left(\int_0^{((y_0 - y)/v)} e^{-y_0 \tau + v \tau^2 / 2} d\tau + 1\right)^2} \left(1 - \frac{e^{\frac{y^2 - y_0^2}{2v}}}{\int_0^{((y_0 - y)/v)} e^{-y_0 \tau + v \tau^2 / 2} d\tau + 1}\right)^{n-1} dy_0. \end{aligned} \quad (26)$$

where an arbitrary upper limit of $(q - 2)s$ is imposed to restrict to regions where the deterministic supply rate of mutants is guaranteed to hold, with contributions from mutants founded when the class was at fitness greater than $(q - 2)s$ already negligible for very rare mutations. The total SFS of semi-private variants is then obtained by integrating over all final fitnesses y :

$$\begin{aligned} F_{rare}(n/N) &= \int_{-qs}^{(q-2)s} F_{rare}(n/N, y) dy \\ &= \frac{1}{v} \frac{U_b N}{\sqrt{2\pi v}} \int_{-\infty}^\infty \int_y^\infty \frac{e^{\frac{y^2 + 2y_0 s - 2y_0^2 - s^2}{2v}}}{\left(\int_0^{((y_0 - y)/v)} e^{-y_0 \tau + v \tau^2 / 2} d\tau + 1\right)^2} \\ &\quad \times \left(1 - \frac{e^{\frac{y^2 - y_0^2}{2v}}}{\int_0^{((y_0 - y)/v)} e^{-y_0 \tau + v \tau^2 / 2} d\tau + 1}\right)^{n-1} dy_0 dy. \end{aligned} \quad (27)$$

To extract the leading order behavior, we consider frequencies just at the mean, $y = 0$:

$$F_{rare}(n/N, 0) = \frac{1}{v} \frac{U_b N}{\sqrt{2\pi v}} \int_0^\infty \frac{e^{\frac{2y_0 s - 2y_0^2 - s^2}{2v}}}{\left(\int_0^{y_0/v} e^{-y_0\tau + v\tau^2/2} d\tau + 1\right)^2} \left(1 - \frac{e^{-\frac{y_0^2}{2v}}}{\int_0^{y_0/v} e^{-y_0\tau + v\tau^2/2} d\tau + 1}\right)^{n-1} dy_0. \quad (28)$$

Examining the integral inside the integrand, we observe that

$$\int_0^{y_0/v} e^{-y_0\tau + v\tau^2/2} d\tau = \sqrt{\frac{2}{v}} e^{-y_1^2} \int_0^{y_1} e^{y^2} dy = \sqrt{\frac{2}{v}} D(y_1) \quad (29)$$

for $y_1 = y_0/\sqrt{2v}$, and $D(y_1)$ is Dawson's integral, a well-studied special function. If y_1 is small, then

$$D(y_1) \approx y_1 - \frac{2}{3}y_1^3 + \frac{4}{15}y_1^5 \dots \quad (30)$$

Since we are considering only small lineage sizes n , the integral will obtain its main contribution for y_0 small (meaning that most nearly-private variants were founded recently in the past). Thus we can use the first order expansion:

$$\begin{aligned} F_{rare}(n/N, 0) &\approx \frac{U_b N}{v\sqrt{2\pi v}} \int_0^\infty \frac{e^{\frac{2y_0 s - 2y_0^2 - s^2}{2v}}}{(y_0/v + 1)^2} \left(1 - \frac{1}{y_0/v + 1}\right)^{n-1} dy_0 \\ &= \frac{U_b N v e^{-s^2/(2v)}}{\sqrt{2\pi v}} \int_0^\infty \exp\left(\frac{y_0 s - y_0^2}{v}\right) \frac{y_0^{n-1}}{(y_0 + v)^{n+1}} dy_0 \\ &\approx \frac{U_b N v e^{-s^2/(2v)}}{\sqrt{2\pi v}} \left(\int_0^\infty \exp\left(\frac{-y_0^2}{v}\right) \frac{y_0^{n-1}}{(y_0 + v)^{n+1}} dy_0 \right. \\ &\quad \left. + \frac{s}{v} \int_0^\infty \exp\left(\frac{-y_0^2}{v}\right) \frac{y_0^n}{(y_0 + v)^{n+1}} dy_0 \right) \\ &= \frac{U_b N e^{-s^2/(2v)}}{\sqrt{2\pi}} \left(\int_0^\infty \frac{e^{-\xi^2} \xi^{n-1}}{(\xi + \sqrt{v})^{n+1}} d\xi + \frac{s}{\sqrt{v}} \int_0^\infty \frac{e^{-\xi^2} \xi^n}{(\xi + \sqrt{v})^{n+1}} d\xi \right), \quad (31) \end{aligned}$$

where, in the last step, the substitution $\xi = \frac{y_0}{\sqrt{v}}$ was performed. Performing each integration separately:

$$\begin{aligned} \int_0^\infty \frac{e^{-\xi^2} \xi^{n-1}}{(\xi + \sqrt{v})^{n+1}} d\xi &\approx \frac{1}{n\sqrt{v}}, \\ \int_0^\infty \frac{e^{-\xi^2} \xi^n}{(\xi + \sqrt{v})^{n+1}} d\xi &\approx -\frac{\gamma}{2} + \log\left(\frac{1}{n\sqrt{v}}\right) + \frac{1}{n} \end{aligned} \quad (32)$$

for $\sqrt{v} \ll 1$. Note that the second approximation breaks down for n large; however, in the

realm of validity of our approximation (derived below), the second term will generally be much smaller than the first term because of the log dependence on n .

Thus, the leading order behavior is

$$F_{rare}(n/N, y = 0) = \frac{U_b N e^{-s^2/(2v)}}{\sqrt{2\pi v n}}. \quad (33)$$

Small n approximation condition: We would like to derive a realm of validity for all of our approximations. The primary assumption made in simplifying the Dawson's integral and exponential integrals is that $y_0/\sqrt{2v} \ll 1$, which is justified if the dominant contribution to the integral occurs for y_0 satisfying this condition. Since we are integrating against a Gaussian, the integral is sharply peaked around the maximum of

$$e^{\frac{2y_0 s - 2y_0^2 - s^2}{2v}} \exp\left(-\frac{(n-1)e^{-\frac{y_0^2}{2v}}}{e^{-y_0^2/2v} \int_0^{y_0/v} e^{vy^2/2} dy + 1}\right). \quad (34)$$

Thus we would like to find the location of the maximum of the exponent's argument:

$$\frac{2y_0 s - 2y_0^2 - s^2}{2v} - \frac{(n-1)e^{-\frac{y_0^2}{2v}}}{e^{-y_0^2/2v} \int_0^{y_0/v} e^{vy^2/2} dy + 1}. \quad (35)$$

Define

$$I = e^{-y_0^2/2v} \int_0^{y_0/v} e^{vy^2/2} dy. \quad (36)$$

The equation to be solved for the peak y_{max} is

$$0 = s - 2y_{max} + \frac{(n-1)y_{max}}{I+1} e^{-y_{max}^2/(2v)} + \frac{(n-1)y_{max} I e^{-y_{max}^2/(2v)}}{(I+1)^2}. \quad (37)$$

Note that $y_{max} = \mathcal{O}(s/2)$, with some n dependent correction that necessarily increases the location of the peak. Thus, for any of these Taylor expansions to hold for any n , we require at minimum that $s \ll \sqrt{v}$.

In this case it is true that

$$y_{max} = \frac{s}{2} + \frac{(n-1)y_{max}}{2(I+1)} e^{-y_{max}^2/(2v)} + \frac{(n-1)y_{max} I e^{-y_{max}^2/(2v)}}{2(I+1)^2} < \frac{s}{2} + \frac{(n-1)y_{max}}{2(I+1)} + \frac{(n-1)y_{max} I}{2(I+1)^2}. \quad (38)$$

This may be solved numerically to find the exact location of the peak. However, if we suppose that the n terms are a small perturbation on the $s/2$ peak, and (as we have already assumed) $s \ll \sqrt{v}$, we are justified in a first order expansion of Dawson's integral:

$$\begin{aligned} y_{max} &< \frac{s}{2} + \frac{(n-1)y_{max}}{2(y_{max}/v+1)} + \frac{(n-1)y_{max}^2/v}{2(y_{max}/v+1)^2} \\ &< \frac{s}{2} + (n-1)v. \end{aligned} \quad (39)$$

So long as $(n-1)v \ll 1$, it is indeed true that the new maximum is a small perturbation around the $s/2$ peak, and our original first order expansion of I was justified.

Thus, for the small n approximation to hold, we require that $y_{max} < \frac{s}{2} + (n-1)v \ll \sqrt{2v}$. Thus, the small n approximation holds for

$$n \ll \sqrt{\frac{2}{v}} - \frac{s}{2v}. \quad (40)$$

Derivation of the Neutral Site Frequency Spectrum:

In this section we derive the asymptotic form of the neutral site frequency spectrum for common alleles. To derive this, we start from the assumption of a class that is growing exponentially (and deterministically), such that

$$n_{k-1}(t) = \frac{e^{(q-1)st}}{qs}. \quad (41)$$

This class supplies neutral mutants at a rate U_n . Thus the expected number of individuals with lineage sizes n at time t , $F(n, t)$, is simply the integral over the expected number of mutants introduced at time τ multiplied by the probability for a mutant to reach a lineage size n in time $t - \tau$:

$$F(n, t) = \int_{-\infty}^t U_n n_{k-1}(\tau) P(n, t - \tau) d\tau. \quad (42)$$

The probability for a lineage to reach size n in a time t is given in Desai and Fisher (2007) and in the first section of this SI. This gives

$$F(n, t) = ((q-1)s)^2 \int_{-\infty}^t \frac{U_n d\tau}{qs} e^{(q-1)s\tau} \left(\frac{(1 + (q-1)s)e^{(q-1)s(t-\tau)} - 1 - (q-1)s}{((1 + (q-1)s)e^{(q-1)s(t-\tau)} - 1)} \right)^n$$

$$\times \frac{e^{(q-1)s(t-\tau)}}{((1+(q-1)s)e^{(q-1)s(t-\tau)} - 1)((1+(q-1)s)e^{(q-1)s(t-\tau)} - 1 - (q-1)s)}. \quad (43)$$

Defining $\sigma = (q-1)s$, $y = (1+\sigma)e^{\sigma k} - 1 - \sigma$, we obtain

$$F(n, t) = \frac{\sigma U_n e^{\sigma t}}{qs} \int_0^\infty \frac{dy y^{n-1}}{(y+1+\sigma)(y+\sigma)^{n+1}}. \quad (44)$$

An explicit series expansion in powers of n may now explicitly be derived. First we note that

$$\frac{1}{y+1+\sigma} = \frac{1}{y+\sigma} \left(\frac{1}{1 + \frac{1}{y+\sigma}} \right) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(y+\sigma)^{k+1}}, \quad (45)$$

which gives

$$F(n, t) = \frac{\sigma U_n e^{\sigma t}}{qs} \int_0^\infty \sum_{k=0}^{\infty} \frac{dy y^{n-1} (-1)^k}{(y+\sigma)^{n+k+1}} = \frac{U_n e^{\sigma t}}{qs} \sum_{k=2}^{\infty} \frac{(-1)^k (k-1)! \Gamma(n)}{\sigma^{k-1} \Gamma(k+n)} = \frac{U_n e^{\sigma t}}{qs} \sum_{k=2}^{\infty} \frac{(-1)^k \beta(k, n)}{\sigma^{k-1}}, \quad (46)$$

where the Beta function $\beta(k, n)$ is defined as

$$\beta(k, n) = \frac{\Gamma(k) \Gamma(n)}{\Gamma(n+k)}. \quad (47)$$

The above is well approximated by the following expansion:

$$F(n, t) = \frac{U_n e^{\sigma t}}{qs} \sum_{k=2}^{\infty} \frac{(-1)^k (k-1)!}{\sigma^{k-1} n^k} \approx \frac{U_n e^{\sigma t}}{qs} \left(\frac{1}{\sigma n^2} - \frac{2}{\sigma^2 n^3} + \frac{6}{\sigma^3 n^4} \dots \right). \quad (48)$$

As expected, for $\sigma \rightarrow \infty$ or $n \rightarrow \infty$ we recover the characteristic n^{-2} decay of an exponentially expanding population. Keeping the leading order term, we obtain

$$F(n, t) = \frac{U_n e^{\sigma t}}{qs \sigma n^2} \Rightarrow f(x) dx = \frac{U_n dx}{\sigma x^2}. \quad (49)$$

Details concerning forward-time Wright Fisher simulations:

We validate some of our results in the text by comparing theorized predictions to simulations. Toward this end, we implemented forward-time simulations that closely resemble evolution in the Wright-Fisher model. The details of the implementation of these simulations

is described in detail in our previous work (Good et al., 2012).

To measure the transition probabilities and sojourn times of mutations between fitness classes, an initially clonal population is allowed to evolve for $2(q + 1)\langle\tau_k\rangle$ generations until it reaches its steady state distribution of fitnesses. At this point, a mutation is seeded in at a frequency $x = 0.5$ in each fitness class. A new class k is allowed to establish, and shortly afterwards the mutation reaches some steady frequency x_k in this class. The population is then allowed to evolve until a class containing $\lceil k + q \rceil$ beneficial mutations establishes. At this point, class k is (roughly) at the population's mean fitness, and the frequency of the mutant in this class is recorded. This prescription certifies that frequencies of mutations attain their long-time steady values long before they are measured. Upon the establishment of class $\lceil k + q + 1 \rceil$, the frequency of the mutant in class $k + 1$ is recorded. Generally, when class $\lceil k + q + i \rceil$ establishes, the frequency of the mutant in class $k + i$ is recorded. In this way, a vector of transitions $\{x_k, x_{k+1}, \dots\}$ is generated, until $x_i = 0$ or $x_i = 1$ in the mean class for some i . The transition and sojourn probabilities are then collected from 60,000 such runs for each parameter set.

Implementation of our code in Python (used to obtain transition and sojourn times) and C (used to obtain site frequency spectra) are freely available upon request.

LITERATURE CITED

- DESAI, M. AND FISHER, D. 2007. Beneficial mutation- selection balance and the effect of linkage on positive selection. Genetics 176:1759.
- DESAI, M., WALCZAK, A., AND FISHER, D. 2013. Genetic diversity and the structure of genealogies in rapidly adapting populations. Genetics 193:565–585.
- GOOD, B. H., ROUZINE, I. M., BALICK, D. J., HALLATSCHKE, O., AND DESAI, M. M. 2012. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. Proc. Natl. Acad. Sci. USA 109:4950–4955.
- KENDALL, D. G. 1948. On the generalized birth-and-death process. Annals of Mathematical Statistics 19:1–15.