Genealogies of rapidly adapting populations

Richard A. Neher[a,1] and Oskar Hallatschek[b]

[a]Evolutionary Dynamics and Biophysics Group, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany; and [b]Biophysics and Evolutionary Dynamics Group, Max Planck Institute for Dynamics and Self-Organization, 37077 Göttingen, Germany

The genetic diversity of a species is shaped by its recent evolutionary history and can be used to infer demographic events or selective sweeps. Most inference methods are based on the null hypothesis that natural selection is a weak or infrequent evolutionary force. However, many species, particularly pathogens, are under continuous pressure to adapt in response to changing environments. A statistical framework for inference from diversity data of such populations is currently lacking. Towards this goal, we explore the properties of genealogies in a model of continual adaptation in asexual populations. We show that lineages trace back to a small pool of highly fit ancestors, in which almost simultaneous coalescence of more than two lineages frequently occurs. Whereas such multiple mergers are unlikely under the neutral coalescent, they create a unique genetic footprint in adapting populations. The site frequency spectrum of derived neutral alleles, for example, is nonmonotonic and has a peak at high frequencies, whereas Tajima's D becomes more and more negative with increasing sample size. Because multiple merger coalescents emerge in many models of rapid adaptation, we argue that they should be considered as a null model for adapting populations.

coalescent theory | demographic inference | pathogen evolution | population genetics

Evolutionary change is usually too slow to be observed in real time. A sequence sample represents a static snapshot from which we want to learn about a dynamic evolutionary process. The predominant framework to analyze such population genetic data and infer demographic history is Kingman's neutral coalescent. Within this model, all individuals are equivalent (i.e., there are no fitness differences), and pairs of lineages merge at random. The statistical properties of genealogies in this simple population genetic model can be computed exactly (1, 2), facilitating comparison with data. One central prediction of the neutral coalescent is that the genetic diversity of a population is proportional to its size. This prediction, however, is at odds with the observed weak correlation between genetic diversity and population size, a paradox often remedied by the definition of an effective population size proportional to the genetic diversity. The model has been generalized to account for historic changes in population size, mutation rates, geographical structure, and effects of purifying selection (3–7). Positive selection, however, has proved difficult to incorporate, and progress has been limited to rare selective sweeps (8, 9) and weak selection (10).

In many populations, particularly large microbial populations, selection is neither rare nor weak. Instead, these populations are under sustained pressure to adapt to changing environments. Prominent examples include pathogens like influenza that continuously evade human immune responses or HIV, which establishes a chronic infection despite heavy immune predation. The genealogical trees reconstructed from sequence samples often suggest substantial departure from neutrality; ref. 11 has examples from viral evolution, and ref. 12 has eukaryotic examples. The influenza tree shown in Fig. 1, for instance, is incompatible with a neutral genealogy, because there are parts where many lineages merge in a very brief period, and the tree often branches extremely unevenly, with very few individuals on one branch and many individuals on the other branch. These two observations represent fundamental deviations from the standard neutral model, even when a varying population size is allowed. Strelkova and Lässig

(14) present a detailed analysis of Influenza A evolution and conclude that influenza is governed by coalescence processes different from the Kingman's coalescent.

To analyze and interpret genealogies of populations under sustained directional selection, an alternative simple null model would be extremely useful. The features of genealogies discussed above are, in fact, common to a class of non-Kingman's coalescence models, which have received considerable attention in the mathematical coalescent literature (15, 16). A special case is the Bolthausen–Sznitman coalescent (BSC) (17), which has been shown to describe the genealogies in models where a population expands into uninhabited territory (18). On the basis of a particular exact solution and a phenomenological theory, Brunet et al. (18) conjectured that genealogies in all models of the same universality class [the class of stochastic Fisher–Kolmogorov–Petrovskii–Piscounov (FKPP) waves] (19, 20) are described by the BSC (recent review in ref. 21). This universality class contains all models with short-range dispersal and logistic growth with constant rate in partially filled demes.

We will argue in this article that the BSC emerges generically in models of rapidly adapting asexual populations in a similar way as it describes genealogies in traveling waves of FKPP type. We present extensive computer simulations and investigate the distribution of heterozygosity in the population, the average time to the most recent common ancestor, and the site frequency spectrum (SFS). Most notably, the SFS is nonmonotonic with a large number of high frequency-derived alleles. We then study a simplified model analytically and show that the underlying genealogical process is approximately the BSC. In the discussion, we outline the basic features of the BSC and discuss its applicability to wider classes of models.

Model

The evolutionary dynamics of a large population are mainly determined by the distribution of fitness in the population. In general, fitness depends on many traits, which are affected by mutations. In a rapidly changing environment, populations are far from any fitness optimum, with many mutations available that increase fitness (and even more that decrease fitness).

To model such scenarios, we consider a collection of N asexual individuals that are characterized by a log-fitness y, which determines their average reproductive success. Specifically, the number offspring of an individual is Poisson-distributed with mean $\exp(y - \lambda)$, where $\lambda = \bar{y} - 1 + N/N_0$ keeps the population size roughly at N_0. The log-fitness of individuals is changed by mutation with probability μ per generation, where the mutational effect, δ, is drawn from a distribution $K(\delta)$. The balance between frequent mutation and selection results in a population that behaves as a traveling pulse along the fitness axis with a steady fitness variance σ^2 (Fig. 2). Absolute fitness itself is, of
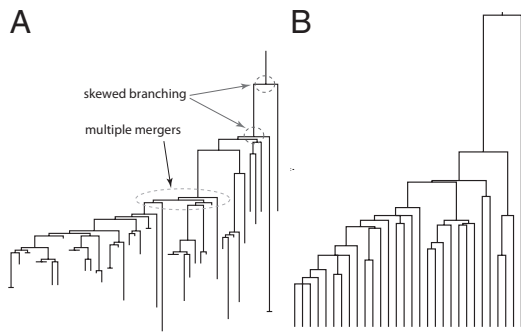
Fig. 1. *A* shows a maximum-likelihood tree of influenza nucleotide sequences (HA segment) sampled in Asia in 2009 (subtype H3N2) produced using Fasttree (13). *B* shows a tree drawn from a simulation of our model of adapting populations. Both trees often branch very unevenly, with almost all descendants on the left-most branch. Although approximate multiple mergers are common in both trees, the influenza tree does not display the uniformly long terminal branches that we observe in simulations. This could be caused by heterogeneous sampling of influenza. Trees are drawn with Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

course, not increasing indefinitely, but increasing fitness is offset by environmental deterioration and deleterious mutations.

We have implemented this model as a computer program (*SI Appendix*) that allows for different mutation distributions $K(\delta)$. In addition, the program keeps track of the parents of each new individual and thereby, saves the complete genealogy of the population. Individuals not leaving any offspring are removed from the genealogical record. From this genealogical record, quantities like pair coalescence times are readily obtained. Furthermore, we can calculate SFSs of neutral mutations by integrating over all positions in the genealogies where such mutations might have occurred.

Similar models have been used by a number of authors (22–27) who have studied the rate of adaptation in these models. Here, we focus on genealogies and their relation to observed genetic diversity. If mutations are frequent relative to the typical effect size of mutations, the model has a continuous time limit described by a stochastic differential equation for the distribution $c(y,t)$ of log-fitness y in the population (26, 28, 29)

$$\partial_t c(y,t) = D\partial_y^2 c(y,t) - \Delta_\mu \partial_y c(y,t) + (y-\lambda)c(y,t) + \text{drift}, \quad [1]$$

where the last term represents the stochastic nature of reproduction (derivation in *SI Appendix*). The diffusion constant and the average mutation input are given by $D = \mu\langle\delta^2\rangle/2$ and $\Delta_\mu = \mu\langle\delta\rangle$, respectively, where the average $\langle\ldots\rangle$ is over the distribution of mutational effects $K(\delta)$. The exact form of the distribution of mutational effects and the relative importance of deleterious and beneficial mutations are irrelevant as long as this diffusive approximation is valid (*SI Appendix*). Unless otherwise stated, we use $\mu=1$ and draw mutational effects from a Gaussian distribution with variance s^2 and zero mean.

In this model, large populations attain a steady fitness distribution of roughly Gaussian shape with variance $\sigma^2 \approx D^{\frac{2}{3}}(24 \log \tilde{N})^{\frac{1}{3}}$, where $\tilde{N} = N D^{\frac{1}{3}}$ (22, 28). The distribution translates to higher fitness with a velocity $v = \sigma^2 + \Delta_\mu$. The distribution and its landmarks are sketched in Fig. 2. It is convenient to measure log-fitness relative to the population mean, $x = y - \bar{y}$. The fittest individuals of the population reside roughly $x_c \approx \sigma^4/4D$ above the population mean. Computer programs and analysis scripts are available on the authors' Web site (http://www.eb.tuebingen.mpg.de/research/research-groups/richard-neher.html).

Results

We first present simulation results of our model and contrast the patterns of genetic diversity of continuously adapting popula-

tions with neutral expectations. Below, we will analyze our model mathematically and show that the striking differences result from the exponential amplification of individual lineages by selection.

Distribution of Heterozygosity and Pair Coalescence Times. Assuming a molecular clock, the expected number of neutral differences between two genomes is $\pi = 2T_2\mu_n$, where μ_n is the neutral mutation rate and T_2 is the time to the most recent ancestor of the pair of sequences. Across many realizations of the process (e.g., independent loci), T_2 follows a distribution $P(T_2)$, which in the neutral case, is exponential with mean N. Simulation results for our model shown in Fig. 3 display a very different distribution of T_2 and equivalently, π. Very few pairs of sequences coalesce early, which results in the long terminal branches observed in trees (Fig. 1B). We then observe a peak in coalescence around $t \approx \sigma^2/2D$, after which the distribution of pair coalescence times decays exponentially with a characteristic time constant proportional to $\sigma^2/2D$. Within a neutral coalescent framework, a distribution of this kind would be interpreted as a rapid population expansion starting $\sigma^2/2D$ in the past. Before this expansion, the population size would be estimated to have been constant at $N_e \propto \sigma^2/2D$. However, the size of the population did not change in our model. Instead, the population was adapting by many small steps, and the conclusion that N increased in the past is wrong.

Two lineages chosen at random from the population are most likely from near the center of the fitness distribution. There are many individuals in this part of the distribution, and therefore, the probability of immediate coalescence is low. Although the sampled individuals are typical, their ancestors tend to have higher than average fitness. Only after ancestral lineages have moved to the high fitness tail of the distribution, where only few individuals are, does the rate of coalescence become appreciable. This migration of lineages to higher fitness is a well-known effect (6, 30, 31), and it is illustrated in Fig. 2. The speed at which lineages move to higher (relative) fitness is initially σ^2 (the speed of the mean minus the mutational input), whereas they slow down as they reach the tip. Consistent with the above interpretation, the delay of coalescence, T_{delay}, is roughly two times the time required for the mean fitness to catch up with the high fitness nose (i.e., $T_{delay} \approx 2x_c\sigma^{-2} = \sigma^2/2D$).
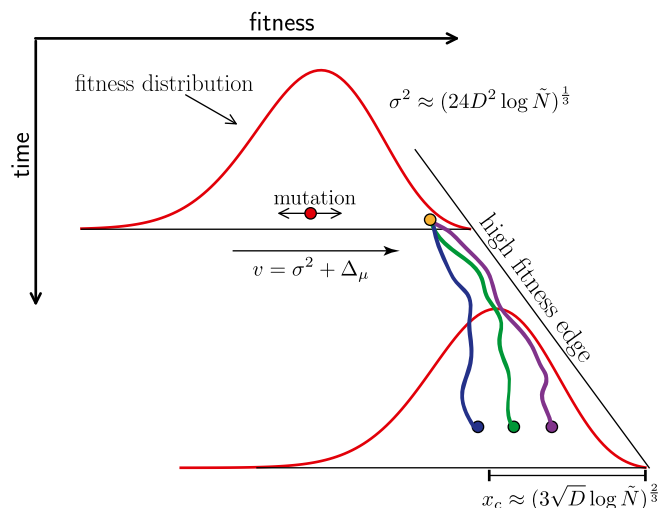
Fig. 2. Ancestral lineages in evolving populations. The figure shows the fitness distribution of the population, translating to higher fitness with velocity $\sigma^2 + \Delta_\mu$, at two time points. Randomly sampled individuals (green, blue, and violet dots in the later population) tend to come from the center of the distribution, whereas ancestors tend to be among the fittest in the population. The ancestral lineages wiggle because of mutations that randomly perturb their fitness. Simultaneously, lineages move to the high fitness edge, where they are likely to meet and coalesce. The fittest individuals are typically at $x_c \approx \sigma^4/4D$ above the mean fitness.
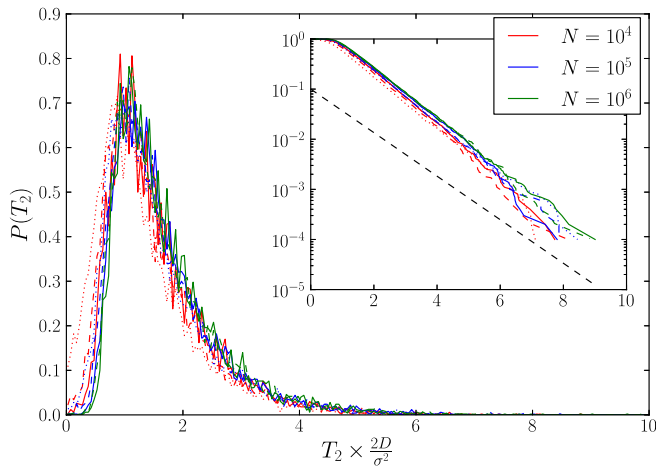
Fig. 3. The distribution of pair coalescence times (proportional to heterozygosity) in a model of rapidly adapting populations. After rescaling time by $\sigma^2/2D$, curves for different N and s collapse onto a single master curve. This collapse shows that $\sigma^2/2D \propto D^{-\frac{1}{3}}(\log \bar{N})^{\frac{1}{3}}$ is the timescale of coalescence. After a delay, $T_{delay} \approx \sigma^2/2D$, T_2 is exponentially distributed, which is apparent from the *Inset* showing the cumulative distribution $P(T_2 > T)$. An exponential $\exp[-2TD/\sigma^2]$ is indicated as a black dashed line. Different line styles correspond to $s = 0.01$ (solid), $s = 0.001$ (dashed), and $s = 0.0001$ (dotted), whereas the mutation rate is $\mu = 1$. For each parameter combination, random pairs are sampled at 10,000 time points $2s^{-2/3}$ generations apart. Fig. S1 shows the corresponding distributions of T_3 and T_4.

After lineages have moved to the high fitness tail, they seem to coalesce uniformly with a time constant $T_c \approx \sigma^2/2D$. From the dependence of σ^2 on population parameters, we see that $T_c \propto (\log \bar{N})^{\frac{1}{3}}$ increases only weakly with the population size.

Site Frequency Spectra. The density $f(\nu)d\nu$ of neutral-derived alleles in the frequency interval $[\nu, \nu + d\nu]$ is known as the SFS. The neutral SFS is a convenient summary of the neutral diversity segregating in the population. A mutation that happened on a particular branch of the genealogy will later be present in all individuals that descend from this branch. Hence, the SFS harbors information about the distribution of branch weights and the branch length of the genealogy. In Kingman's coalescent, the SFS is simply given by $f(\nu) = \Theta/\nu$, where $\Theta = 2\langle T_2 \rangle \mu$ is the average heterozygosity. Importantly, it is a monotonically decreasing function of the frequency. Fig. 4 shows SFSs measured in simulations of our model. The most striking qualitative difference is the nonmonotonicity, a feature known to be common in the presence of selective sweeps caused by hitchhiking (32).

The nonmonotonicity of $f(\nu)$ implies the existence of long branches deep in the tree that are ancestors of almost everybody in the population, whereas a small minority of the population descends from different lineages. Such very asymmetric branchings are unlikely in Kingman's coalescent, where at any split, the fraction of individuals that go left or right is uniformly distributed (2). Such asymmetric branchings are common in our model and frequently observed in reconstructed genealogies from rapidly adapting organisms (Fig. 1*A*).

The axes in Fig. 4 are scaled to facilitate the comparison with analytic results. At low frequencies, the SFS is proportional to ν^{-2} (33, 34) and hence much steeper than the neutral SFS in Kingman's coalescent, $f(\nu) \propto \nu^{-1}$. Hence, samples will be dominated by singletons. In addition, $f(\nu)$ is nonmonotonic and increases as $\nu \to 1$.

The majority of the contributions to the increase of $f(\nu)$ for $\nu \to 1$ stems from the very last coalescent event. In this last coalescent event, two or more lineages are merging. One of these lineages is typically the ancestor of almost the entire sample, whereas the others share the remaining minority. The distribution of the offspring of these lineages and their number has been studied by

Goldschmidt and Martin (35), who showed that the distribution of the size of the biggest lineage is asymptotically $\propto (1 - \nu)^{-1}$. In *SI Appendix*, section III, we derive the more accurate approximation

$$f(\nu) \approx \frac{T_c \mu}{(\nu - 1)\log(1 - \nu)} \quad \text{for} \quad 1 - \nu \ll 1. \qquad [2]$$

To compare the SFS of our model with SFS of the BSC across the entire range of ν, we simulated the idealized BSC and find very good agreement (solid black line in Fig. 4). The SFS of the idealized BSC deviates from the SFS of the model of adaptation only at very low allele frequencies. The model of adaptation tends to have even more rare alleles than the BSC, which is because of the fact that lineages have to move to the high fitness tail before coalescence begins.

The nonmonotonicity $f(\nu)$ is a clear indication that the genealogies in this model with selection are fundamentally different from canonical neutral genealogies (Kingman). In Kingman's coalescent, neither constant nor exponentially growing population sizes gives rise to nonmonotonic SFS (Fig. S2).

Time to the Most Recent Common Ancestor. In Kingman's coalescent, the expected time to the most recent common ancestor (MRCA) of a sample of size n, $\langle T_{MRCA} \rangle = N(2 - 2/n)$, increases only very slowly with n. This saturation is a consequence of the even branching ratios; an additional individual will most likely coalesce with existing samples and only rarely increase T_{MRCA}. In contrast, the trees generated by our model of adaptation tend to branch very unevenly, and one often observes that one external branch goes all of the way back to the MRCA of the sample, as in Fig. 1*B*. As the sample size is increased, one continues to sample deeper into the tree. This increasing tree depth is a generic property of the BSC (16), where the average $\langle T_{MRCA} \rangle$ increases as log log n with the sample size n. Similarly, $\langle T_{MRCA} \rangle$ of the entire population is expected to increase as T_c log log N with the population size. Our simulations are consistent with this behavior (Fig. S3).

Note that T_c depends weakly on N in adapting populations, whereas it increases linearly with N in Kingman's coalescent. In contrast, the rescaled time to the MRCA, $T_c^{-1} \langle T_{MRCA} \rangle$, asymptotes to two in Kingman's coalescent, whereas it continues to increase with N in the BSC.
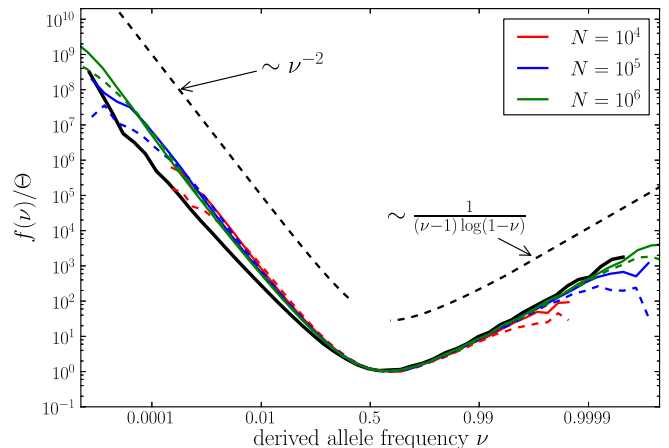
Fig. 4. SFS of (derived) neutral alleles in rapidly adapting populations is nonmonotonic, with peaks at low frequencies and near fixation. The asymptotic behavior of the SFS at low and high derived frequencies is shown as dashed black lines. The solid black line is the SFS of the BSC simulated using Eqs. **7** and **8** with $N = 100{,}000$ averaged over 10,000 runs. Line styles and parameters are as in Fig. 3.

Analysis. The simulation results presented above show that genealogies arising in our model are distinct from those genealogies expected in Kingman's coalescent and display a number of features reminiscent of the BSC. We will now describe how this coalescent process emerges from the dynamics of the model.

Individuals in our model have a heritable fitness that determines the distribution of the number of immediate offspring. Although fit individuals have, on average, more offspring than less-fit individuals, the fitness differences in the population are small, and the offspring distribution across the population is narrow. However, fitness is heritable, and fit individuals can have a very large number of distant great^t-grandchildren. Hence, the distribution of offspring after t generations, $P(n,t)$, will be dominated by fit individuals and can have a very long tail. Conversely, the present-day population has fewer and fewer ancestors as we trace its lineages back in time. At T_{MRCA} generations in the past, there is exactly one individual that is the ancestor to the entire population. Ancestors of the MRCA are also common ancestors (CAs) of the entire population, albeit not the most recent one. Fig. 5 shows that MRCAs and CAs tend to come from the high fitness tail of the population. MRCAs tend to be fitter than CAs, because they are conditioned on giving rise to at least two lineages that persist to the present.

The offspring distribution, $P(n,t)$, changes slowly from the initial narrow distribution to a broad distribution with a power law at intermediate times (34). The broad distribution at intermediate times is at the heart of the correspondence of genealogies in models of adapting populations and the BSC.

The BSC assumes that all individuals are exchangeable and that, in every coalescent event, a randomly chosen set of lineages merges into one. Each possible merger event has a specific rate associated with it, and the rate at which k individuals merge into one CA is $q_k = T_2^{-1}(k-1)^{-1}$ (16) (the general expression for the rates is given below in Eq. **7**). In contrast, in the neutral coalescent, higher-order coalescence is very rare, $\propto N^{-(k-1)}$. We will present the basic properties of the BSC briefly in the discussion.

To appreciate how these coalescence rates can emerge from a model with selection, consider the number of individuals n_i that descend from an individual i that lived t generations in the past. The probability that k individuals sampled randomly from the population have a CA t in the past is then given by (Eq. **3**)

$$Q_k(t) = \left\langle \sum_{i=1}^{N} \left(\frac{n_i}{\sum_j n_j} \right)^k \right\rangle, \quad \text{[3]}$$

where the average $\langle . \rangle$ is over all n_i. $Q_k(t)$ is dominated by $n_i \gg k$, and therefore, sampling with replacement in Eq. **3** is an accurate

approximation. Using the identity $\Gamma(k)C^{-k} = \int_0^z dz\, z^{k-1}e^{-zC}$ and assuming that t is small enough that the different n_i values are still approximately independent, we can express Q_k as

$$Q_k(t) \approx \frac{N}{\Gamma(k)} \int_0^\infty dz\, z^{k-1} \langle e^{-zn} \rangle^{N-1} \langle n^k e^{-zn} \rangle$$

$$\approx -\frac{N}{\Gamma(k)} \int_0^\infty dz\, z^{k-1} e^{-N\Phi_z(t)}(-1)^k \partial_z^k \Phi_z(t), \quad \text{[4]}$$

where we introduced the Laplace transform $1 - \Phi_z(t) = \sum_n e^{-zn} P(n,t)$ and assumed $N\Phi_z^2(t) \ll 1$. In *SI Appendix*, we show that $\Phi(t) \sim z^{\frac{\sigma^2}{2Dt}}$ for $t > T_{delay} = \sigma^2/2D$. For a limited interval after $t > T_{delay}$, we find that the probability that k individuals have a CA increases with rate

$$q_k = \frac{2D}{\sigma^2} \frac{1}{k-1} \quad \text{[5]}$$

per unit time. More general coalescence rates can be calculated analogously (*SI Appendix*). Before T_{delay}, the rate of coalescence is very low. This result is in agreement with Fig. 3, where we found that little coalescence happened early, whereas coalescence times are exponentially distributed after that time, with characteristic time $T_c \approx \sigma^2/2D$ for $t > T_{delay}$. The relative rates of mergers of two, three, etc. are consistent with the BSC, explaining our observations for the frequency spectrum and the time to the MRCA.

The branching process approximation used to derive the result in Eq. **5** is valid only for short times but nevertheless, gives us the relative rates of multiple mergers after coalescence begins. For subsequent deeper coalescent events, the relevant lineages are already at the tip of the fitness distribution, and this process repeats itself without the delay. In fact, after this delay, all remaining lineages are in a narrow region at the tip of the fitness distribution. The situation now resembles the situation of coalescence in FKPP waves: The fitness of the lineages is roughly equal, but lineages have to stay ahead of a fitness cutoff to survive. We can, therefore, use the phenomenological theory of genealogies in FKPP waves from ref. 36, which confirms the above result for the coalescent timescale (*SI Appendix*). In *SI Appendix*, we present an additional argument based on tuned models introduced in ref. 26.
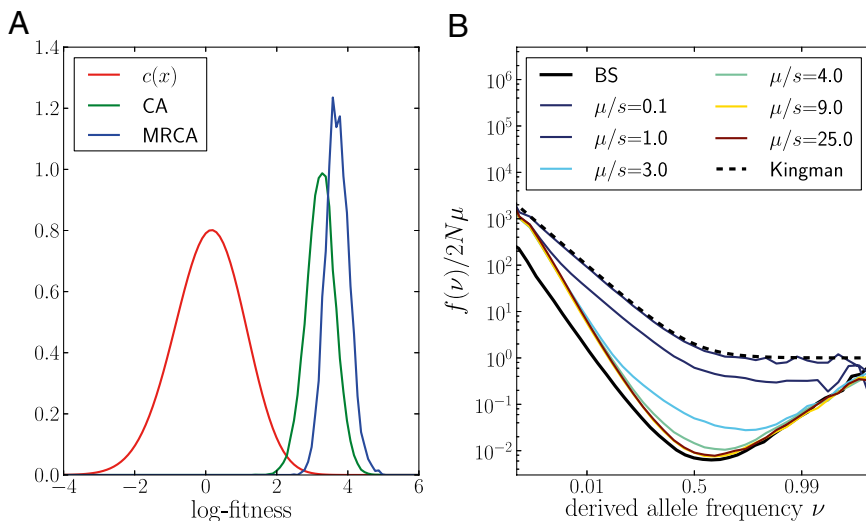
Fig. 5. *A* shows the distribution of the log-fitness of all CAs and all MRCAs compared with the average distribution, $c(x)$, of log-fitness in the population (described in the text). These distributions were measured in forward simulations with $\mu = 1, s = 0.01$, and $N = 10^6$. (*B*) SFS of derived neutral alleles in a background selection scenario with deleterious mutations of effect s. As the ratio μ/s is varied while keeping $\sigma^2 = \mu s = 0.01$ constant, the SFS interpolates between the expectation for the Kingman's coalescent and the BSC ($N = 10^4$).

To corroborate our analysis, we performed additional simulations that allow us to measure the Laplace transform of the distribution of pair coalescent times for very large populations. These simulations show that the pair coalescent time is, indeed, exponential with characteristic time $\sigma^2/2D$ after a delay of the same length (Fig. S4). The algorithm used is similar in spirit to the algorithm by Brunet et al. (18) (*SI Appendix*).

Strictly speaking, the analogy to an exchangeable coalescent model like the BSC requires that different coalescence events that one lineage undergoes be independent. For this finding to be true, individuals descending from a lineage have to distribute evenly across the fitness distribution $c(x)$ between coalescence events, which requires a time $T_{eq} \approx T_c$. Hence, we should not expect a clean convergence to the BSC. Nevertheless, we find it to be a very good model for the observed genealogies after accounting for the delay. The underlying reason is that local equilibration in the region where the ancestral lineages are is fast ($t \approx D^{-1/3}$). This region, however, undergoes fluctuations on the timescale T_{eq} that modulate the overall rate of coalescence but do not significantly affect the local dynamics. For waves of FKPP type that describe the spread of individuals in space, $T_{eq} \ll T_c$ in large populations (18).

Discussion

We have shown that, in a simple model of adapting populations, the observed genealogies are inconsistent with the standard neutral coalescent. Instead, genealogical trees are characterized by long terminal branches and almost simultaneous coalescence of multiple lineages. At branching events deep in the tree, one commonly observes that almost all individuals of the population descend from one branch, whereas very few descend from the other branches. Such skewed branching is unlikely in standard neutral coalescent models, regardless of the history of the effective population size. One consequence of these uneven branching ratios is a nonmonotonic SFS of derived neutral alleles. Compared with the Kingman's coalescent, the low-frequency part of the SFS is much steeper, whereas the high-frequency part shows a characteristic upturn (Fig. 4).

A given pair of lineages is unlikely to coalesce in the bulk of the fitness distribution. Typically, both lineages move into the high-fitness tip of the population distribution before they coalesce as illustrated in Fig. 2. These dynamics result in long terminal branches and a distribution of heterozygosities peaked at intermediate values. After this delay, the typical time to coalescence is again on the order of the time that it takes the fittest individuals to dominate the population (Fig. 3). In panmictic populations, this time depends on the logarithm of the population size, and in our model, it is proportional to $(\log \tilde{N})^3$.

We argue that the exponential amplification of fit lineages is responsible for these observations and that coalescence in such rapidly adapting populations is generically described by a modified BSC (16, 17). The BSC is a special case of the large class of Λ-coalescent processes (15). Given the distribution $p(f)$ of the fraction f of the population that descends from a single individual in the previous generation, the rate at which k of b lineages merges is given by

$$\lambda_{b,k} = \int df\, p(f) f^k (1-f)^{b-k}. \qquad [6]$$

The BSC corresponds to $p(f) \sim f^{-2}$ for large f, in which case Eq. 6 reduces to

$$\lambda_{b,k} = \frac{1}{T_c} \frac{(k-2)!(b-k)!}{(b-1)!} \qquad [7]$$

or Eq. 5 for the special case $b = k$. The total rate at which coalescence events happen in a sample of k lineages is, therefore

$$\lambda_b = \sum_k \binom{b}{k} \lambda_{b,k} = \frac{b-1}{T_c} \qquad [8]$$

in contrast to the neutral coalescent, where $\lambda_b \propto b(b-1)$. A coalescence event reduces the number of surviving lineages on average by $\log b$, and therefore, the average rate at which the number of lineages decreases is $\approx T_c^{-1} b \log b$. The typical time needed to reach the CA of a sample of size n is $\approx T_c \log \log n$, in contrast to $2T_c$ in Kingman's coalescent. The BSC occupies a special intermediate position between Kingman's coalescent, where only pairwise mergers are allowed, and a star coalescent, where all lineages coalesce simultaneously. Star-like genealogies are expected in rapidly expanding populations or a region fully linked to a recent rapid hard sweep (37). In the BSC, multiple mergers (subsets of lineages with star-like trees) are frequent, but at the same time, there are many mergers at different depths in the tree. In fact, the BSC is the $\alpha = 1$ case of the one-parameter family of β-coalescents with parameter $0 < \alpha < 2$, whereas the case $\alpha \to 2$ corresponds to the Kingman's coalescent. A more in-depth discussion is in the recent review by Berestycki (16). The BSC is easily implemented as a computer simulation by drawing an exponentially distributed random number with mean λ_b^{-1} to determine the time of the next event. The type of event is then chosen with probabilities proportional to $\lambda_{b,k}$.

The models of adaptation that we have studied have a narrow offspring distribution. Nevertheless, the exponential amplification of fit genotypes over many generations gives rise to a distribution of clone sizes with the required asymptotic behavior. The important lineages are those lineages that run ahead of the distribution, expand faster, and take over a significant fraction of the population (36). Over even longer times, the fitness of ancestors and descendants decorrelates. This gradual decorrelation allows us to approximate the genealogies with the abstract BSC, which assumes that there are no correlations in offspring number across generations.

Conventionally, an increased variance in offspring number is accounted for by defining an effective population size. With a clone size distribution $p(f) \sim f^{-2}$, however, the variance diverges with the population size (34). Similar effects arise in other models with very skewed offspring distribution (38). As a consequence, the genealogies are dominated by rare anomalously large clones and described by the BSC rather than Kingman's coalescent. The rate of coalescence is not set by N^{-1} but by the rate at which clones expand and collapse. We would like to stress that evolutionary dynamics thereby remain highly stochastic, even in very large populations. Analogous behavior has recently been observed in models of individuals invading uninhabited territory (FKPP type waves) (18) and ensembles of supercritical branching processes (39).

The BSC is not only a good model for genealogies of adapting asexual populations, but also, it applies to populations under purifying selection in which Muller's ratchet clicks often. The standard model for the distortion of genealogies by purifying selection assumes that deleterious mutations are rapidly purged and coalescence is neutral in the mutation-free class with a reduced population size $Ne^{-\mu/s}$, where μ is the deleterious mutation rate and s is the effect size of deleterious mutations (4). More elaborate analysis based on a fitness class coalescent explicitly tracks lineages through the population and calculates the contribution to coalescence before lineages reach the mutation-free class (5). However, this standard model of purifying selection only applies if the mutation-free class is large and Muller's ratchet does not operate, which requires $Nse^{-\mu/s} \ll 1$ (40–42). Fig. 5B shows the SFS of derived neutral alleles for different ratios μ/s. For small μ/s, the SFS is similar to the SFS of Kingman's coalescent with a reduced time to coalescence, in accordance with the background selection theory. However, as soon as the ratchet starts to click frequently, the SFS develops the nonmonotonicity characteristic of the BSC.

If the ratchet is clicking fast, the fitness distribution in the population resembles the distribution of traveling wave models,

but selection on fitness variation cannot keep up with the influx of deleterious mutations. Similarly, populations in a steady balance between deleterious and beneficial mutations (43) have genealogies as found here for rapidly adapting populations. The reason for the qualitative difference in the ratchet regime is the fact that the nose of the wave is not steady but constantly turning over. Different lineages are struggling to get ahead of everybody else and in the frame of reference of the population (that is, relative to mean fitness), exponentially amplified. In contrast, dynamics of lineages in the mutation-free class are neutral if Muller's ratchet does not operate.

In *SI Appendix*, we show that the argument that gave rise to the particular coalescence rates in Eq. **5** can be extended to a large class of models that are controlled by a small and fluctuating population of highly fit individuals. We argue that the BSC generically emerges as a consequence of the exponential amplification of the clones descending from these highly fit individuals together with the seeding of novel lineages. The latter could happen by lucky diffusion to high fitness (our model), large-effect beneficial mutations, or lucky outcrossing. After some time, the distribution of lineage size follows a power law with an exponent close to −2 (23, 34, 44). Given an effective offspring distribution of this shape, the BSC follows (18, 39). In ref. 45, the authors study a model where the mutation rate is much smaller than the typical effect sizes of mutations. They show that, also in this case, the genealogies are well-approximated by the BSC after a delay (ref. 45; see also Figs. S5 and S6). Whether the BSC also describes genealogies in scenarios where fitness is increased in rather large increments (compared with the population diversity) (46, 47) remains an interesting topic for future work.

The compatibility of a sample with the neutral coalescent model is typically assessed using statistics such as Tajima's D (48). Tajima's D compares the average number of pairwise differences with the total number of segregating sites in the sample. In the case of the BSC, the average pairwise diversity is proportional to $\langle T_2 \rangle$, whereas the total number of segregating sites is proportional to $\langle T_2 \rangle n / \log n$ (compared with $\langle T_2 \rangle \log n$ for the Kingman's coalescent). This tremendous excess of segregating sites is a consequence of the very steep SFS at small frequencies and results in $D \propto -\log(n)$.

Sexual populations and recurrent selective sweeps at linked loci can also give rise to multiple mergers in the genealogies (9). However, recombination and sexual reproduction will reduce the effects of linked selection and decouple the genealogies of different loci. Hence, we expect that the coalescent behavior crosses over to Kingman's coalescent as the recombination rate increases—at least in models of panmictic populations. This crossover is observed in models of facultatively sexual populations (34).

Given the apparent universality of the BSC in spatially expanding populations and panmictic adapting populations, it should be included as a prior in popular population genetic and phylogenetic inference programs such as BEAST (49).

1. Kingman J (1982) On the genealogy of large populations. *J Appl Probab* 19A:27–43.
2. Derrida B, Peliti L (1991) Evolution in a flat fitness landscape. *Bull Math Biol* 53(3):355–382.
3. Nordborg M (1997) Structured coalescent processes on different time scales. *Genetics* 146(4):1501–1514.
4. Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4):1289–1303.
5. Walczak AM, Nicolaisen LE, Plotkin JB, Desai MM (2012) The structure of genealogies in the presence of purifying selection: A fitness-class coalescent. *Genetics* 190(2):753–779.
6. O'Fallon BD, Seger J, Adler FR (2010) A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. *Mol Biol Evol* 27(5):1162–1172.
7. Barton NH, Etheridge AM (2004) The effect of selection on genealogies. *Genetics* 166 (2):1115–1131.
8. Barton N (1998) The effect of hitch-hiking on neutral genealogies. *Genet Res* 72(2):123–133.
9. Durrett R, Schweinsberg J (2005) A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Process Appl* 115(10):1628–1657.
10. Krone SM, Neuhauser C (1997) Ancestral processes with selection. *Theor Popul Biol* 51(3):210–237.
11. Bedford T, Cobey S, Pascual M (2011) Strength and tempo of selection revealed in viral gene genealogies. *BMC Evol Biol* 11:220.
12. Seger J, et al. (2010) Gene genealogies strongly distorted by weakly interfering mutations in constant environments. *Genetics* 184(2):529–545.
13. Price MN, Dehal PS, Arkin AP (2009) FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 26(7):1641–1650.
14. Strelkowa N, Lässig M (2012) Clonal interference in the evolution of influenza. *Genetics* 192(2):671–682.
15. Pitman J (1999) Coalescents with multiple collisions. *Ann Probab* 27(4):1870–1902.
16. Berestycki N (2009) Recent progress in coalescent theory. arXiv:math.PR/0909.3985.
17. Bolthausen E, Sznitman A-S (1998) On Ruelle's probability cascades and an abstract cavity method. *Commun Math Phys* 197(2):247–276.
18. Brunet E, Derrida B, Mueller AH, Munier S (2007) Effect of selection on ancestry: An exactly soluble case and its phenomenological generalization. *Phys Rev E Stat Nonlin Soft Matter Phys* 76(4 Pt 1):041104.
19. Fisher RA (1937) The wave of advance of advantageous genes. *Ann Eugen* 7(4):355–369.
20. Kolmogorov A, Petrovskii I, Piscounov N (1937) Etude de l'equation de la diffusion avec croissance de la quantite de matiere et son application a un probleme biologique. *Bull Moscow Univ Math Mech* 1:1–25.
21. Brunet É, Derrida B (2012) Genealogies in simple models of evolution. arXiv:q-bio.PE.
22. Tsimring LS, Levine H, Kessler DA (1996) RNA virus evolution via a fitness-space model. *Phys Rev Lett* 76(23):4440–4443.
23. Desai MM, Fisher DS (2007) Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics* 176(3):1759–1798.
24. Rouzine IM, Wakeley J, Coffin JM (2003) The solitary wave of asexual evolution. *Proc Natl Acad Sci USA* 100(2):587–592.
25. Neher RA, Shraiman BI, Fisher DS (2010) Rate of adaptation in large sexual populations. *Genetics* 184(2):467–481.
26. Hallatschek O (2011) The noisy edge of traveling waves. *Proc Natl Acad Sci USA* 108 (5):1783–1787.
27. Park SC, Krug J (2007) Clonal interference in large populations. *Proc Natl Acad Sci USA* 104(46):18135–18140.
28. Cohen E, Kessler DA, Levine H (2005) Front propagation up a reaction rate gradient. *Phys Rev E Stat Nonlin Soft Matter Phys* 72(6 Pt 2):066126.
29. Good BH, Rouzine IM, Balick DJ, Hallatschek O, Desai MM (2012) Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *Proc Natl Acad Sci USA* 109(13):4950–4955.
30. Hermisson J, Redner O, Wagner H, Baake E (2002) Mutation-selection balance: Ancestry, load, and maximum principle. *Theor Popul Biol* 62(1):9–46.
31. Rouzine IM, Coffin JM (2007) Highly fit ancestors of a partly sexual haploid population. *Theor Popul Biol* 71(2):239–250.
32. Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155(3):1405–1413.
33. Basdevant A-L, Goldschmidt C (2008) Asymptotics of the allele frequency spectrum associated with the Bolthausen-Sznitman coalescent. *Electron J Probab* 13(17):486–512.
34. Neher RA, Shraiman BI (2011) Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics* 188(4):975–996.
35. Goldschmidt C, Martin JB (2005) Random recursive trees and the Bolthausen-Sznitman coalescent. *Electron J Probab* 10(21):718–745.
36. Brunet E, Derrida B, Mueller AH, Munier S (2006) Phenomenological theory giving the full statistics of the position of fluctuating pulled fronts. *Phys Rev E Stat Nonlin Soft Matter Phys* 73(5 Pt 2):056126.
37. Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129(2):555–562.
38. Eldon B, Wakeley J (2006) Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* 172(4):2621–2633.
39. Schweinsberg J (2003) Coalescent processes obtained from supercritical Galton-Watson processes. *Stochastic Process Appl* 106(1):107–139.
40. Stephan W, Chao L, Smale JG (1993) The advance of Muller's ratchet in a haploid asexual population: Approximate solutions based on diffusion theory. *Genet Res* 61(3):225–231.
41. Jain K (2008) Loss of least-loaded class in asexual populations due to drift and epistasis. *Genetics* 179(4):2125–2134.
42. Neher RA, Shraiman BI (2012) Fluctuations of fitness distributions and the rate of Muller's ratchet. *Genetics* 191(4):1283–1293.
43. Goyal S, et al. (2012) Dynamic mutation-selection balance as an evolutionary attractor. *Genetics* 191(4):1309–1319.
44. Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philos Trans R Soc Lond B Biol Sci* 213:21–87.
45. Desai MM, Walczak AM, Fisher DS (2012) Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics*, 10.1534/genetics.112.147157.
46. Gerrish PJ, Lenski RE (1998) The fate of competing beneficial mutations in an asexual population. *Genetica* 102-103(1-6):127–144.
47. Schiffels S, Szöllosi GJ, Mustonen V, Lässig M (2011) Emergent neutrality in adaptive asexual evolution. *Genetics* 189(4):1361–1375.
48. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.
49. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214.