Genetic Variability Under the Seedbank Coalescent

Jochen Blath, Adrián González Casanova, Bjarki Eldon,[1] Noemi Kurt, and Maite Wilke-Berenguer
TU Berlin, Institut für Mathematik, 10623 Berlin, Germany

ABSTRACT We analyze patterns of genetic variability of populations in the presence of a large seedbank with the help of a new coalescent structure called the seedbank coalescent. This ancestral process appears naturally as a scaling limit of the genealogy of large populations that sustain seedbanks, if the seedbank size and individual dormancy times are of the same order as those of the active population. Mutations appear as Poisson processes on the active lineages and potentially at reduced rate also on the dormant lineages. The presence of "dormant" lineages leads to qualitatively altered times to the most recent common ancestor and nonclassical patterns of genetic diversity. To illustrate this we provide a Wright–Fisher model with a seedbank component and mutation, motivated from recent models of microbial dormancy, whose genealogy can be described by the seedbank coalescent. Based on our coalescent model, we derive recursions for the expectation and variance of the time to most recent common ancestor, number of segregating sites, pairwise differences, and singletons. Estimates (obtained by simulations) of the distributions of commonly employed distance statistics, in the presence and absence of a seedbank, are compared. The effect of a seedbank on the expected site-frequency spectrum is also investigated using simulations. Our results indicate that the presence of a large seedbank considerably alters the distribution of some distance statistics, as well as the site-frequency spectrum. Thus, one should be able to detect from genetic data the presence of a large seedbank in natural populations.

KEYWORDS Wright–Fisher model; seedbank coalescent; dormancy; site-frequency spectrum; distance statistics

MANY microorganisms can enter reversible dormant states of low [respectively (resp.) zero] metabolic activity, for example when faced with unfavorable environmental conditions; see, *e.g.*, Lennon and Jones (2011) for a recent overview of this phenomenon. Such dormant forms may stay inactive for extended periods of time and thus create a seedbank that should significantly affect the interplay of evolutionary forces driving the genetic variability of the microbial population. In fact, in many ecosystems, the percentage of dormant cells compared to the total population size is substantial and sometimes even dominant (for example, ~20% in human gut, 40% in marine water, and 80% in soil; *cf.* Lennon and Jones 2011, box 1, table *a*). This abundance of dormant forms, which can be short-lived as well as stay inactive for significant periods of time (decades- or century-old spores are not uncommon), thus creates a seedbank that buffers against environmental change, but potentially also against

classical evolutionary forces such as genetic drift, mutation, and selection.

In this article, we investigate the effect of large seedbanks (that is, comparable to the size of the active population) on the patterns of genetic variability in populations over macroscopic timescales. In particular, we extend a recently introduced mathematical model for the ancestral relationships in a Wright–Fisherian population of size N with geometric seedbank age distribution (*cf.* Blath *et al.* 2015b) to accommodate different mutation rates for "active" and "dormant" individuals, as well as a positive death rate in the seedbank. The resulting genealogy, measured over timescales of order N, can then be described by a new universal coalescent structure, the "seedbank coalescent with mutation," if the individual initiation and resuscitation rates between active and dormant states as well as the individual mutation rates are of order $1/N$. Measuring times in units of N and mutation rates in units of $1/N$ is of course the classical scaling regime in population genetic modeling; in particular, the classical Wright–Fisher model has a genealogy that converges in precisely this setup to the usual Kingman coalescent with mutation (Kingman 1982a,b,c; see Wakeley 2009 for an overview).

We provide a precise description of these (seedbank) coalescents and corresponding population models, in part

motivated by recent research in microbial dormancy (Jones and Lennon 2010; Lennon and Jones 2011), in the next section. We argue that our seedbank coalescent is universal in the sense that it is robust to the specifics of the associated population model, as long as certain basic features are captured.

Our explicit seedbank coalescent model then allows us to derive expressions for several important population genetic quantities. In particular, we provide recursions for the expectation (and variance) of the time to the most recent common ancestor (T_{MRCA}), the total number of segregating sites, average pairwise differences, and number of singletons in a sample (under the infinitely many sites model assumptions). We then use these recursions, and additional simulations based on the seedbank coalescent with mutation, to analyze Tajima's D and related distance statistics in the presence of seedbanks and also the observed site-frequency spectrum.

We hope that this basic analysis triggers further research on the effect of seedbanks in population genetics, for example concerning statistical methods that allow one to infer the presence and size of seedbanks from data, to allow model selection [e.g., seedbank coalescent vs. (time-changed) Kingman coalescent], and finally to estimate evolutionary parameters such as the mutation rate in dormant individuals or the inactivation and reactivation rates between the dormant and active states.

It is important to note that our approach is different from a previously introduced mathematical seedbank model in Kaj et al. (2001). There, the authors consider a population of constant size N where each individual chooses its parent a random amount of generations in the past and copies its genetic type from there. The number of generations that separate each parent and offspring can be interpreted as the time (in generations) that the offspring stays dormant. The authors show that if the maximal time spent in the seedbank is restricted to finitely many $\{1, 2, \ldots, m\}$, where m is fixed, then the ancestral process induced by the seedbank model converges, after the usual scaling of time by a factor N, to a time-changed (delayed) Kingman coalescent. Thus, typical patterns of genetic diversity, in particular the normalized site-frequency spectrum, will stay (qualitatively) unchanged. Of course, the point here is that the expected seedbank age distribution is not on the order of N, but uniformly bounded by m, so that for the coalescent approximation to hold one necessarily needs that m is *small* compared to N, which results in a "weak" seedbank effect. This model has been applied in Tellier et al. (2011) in the analysis of seedbanks in certain species of wild tomatoes. A related model was considered in Vitalis et al. (2004), which shares the feature that the time spent in the seedbank is bounded by a fixed number independent of the population size. For a more detailed mathematical discussion of such models, including previous work in Blath et al. (2015a), see Blath et al. (2015b). The choice of the adequate coalescent model [seedbank coalescent vs. (time-changed) Kingman coalescent] will thus also be an important question for study design, and the development of corresponding model selection rules will be part of future research.

Coalescent Models and Seedbanks

Before we discuss the seedbank coalescent, we briefly recall the classical Kingman coalescent for reference—this will ease the comparison of the underlying assumptions of both models.

The Kingman coalescent with mutation

The Kingman coalescent (Kingman, 1982a,b,c) describes the ancestral process of a large class of neutral exchangeable population models including the Wright–Fisher model (Fisher 1930; Wright 1931), the Moran model (Moran 1958), and many Cannings models (Cannings 1974). See, e.g., Wakeley (2009) for an overview. If we trace the ancestral lines (that is, the sequence of genetic ancestors at a locus) of a sample of size n backward in time, we obtain a binary tree, in which we see pairwise coalescences of branches until the most recent common ancestor is reached. Kingman proved that the probability law of this random tree can be described as follows: Each pair of lineages [there are $\binom{n}{2}$ many] has the same chance to coalesce, and the successive coalescence times are exponentially distributed with parameters $\binom{n}{2}, \binom{n-1}{2}, \ldots, 1$, until the last remaining pair of lines has coalesced. This elegant structure allows one to easily determine the expected time to the most recent common ancestor of a sample of size n, which is well known to be

$$\mathbb{E}_n[T_{\text{MRCA}}] = 2\left(1 - \frac{1}{n}\right). \tag{1}$$

Not surprisingly, we will essentially recover (1) for the seedbank coalescent defined below if the relative seedbank size becomes small compared to the active population size.

As usual, mutations are placed upon the resulting coalescent tree according to a Poisson process with rate $\theta/2$, for some appropriate $\theta > 0$, so that the expected number of mutations of a sample of size 2 is just θ.

The underlying assumptions about the population for a Kingman coalescent approximation of its genealogy to be justified are simple but far reaching, namely that the different genetic types in the population are selectively neutral (i.e., do not exhibit significant fitness differences) and that the population size of the underlying population is essentially constant in time. If the population can be described by the (haploid) Wright–Fisher model (of constant size, say N), then, to arrive at the described limiting genealogy, it is standard to measure time in units of N, the *coalescent timescale*, and to assume that the individual mutation rates per generation are of order $\theta/(2N)$. The exact time scaling usually depends on the reproductive mechanism and other particularities of the underlying model (it differs already among variants of the Moran model), but the Kingman coalescent is still a universally valid limit for many *a priori* different

population models (including, *e.g.*, all reproductive mechanisms with bounded offspring variance, dioecy, age structure, partial selfing, and to some degree geographic structure), when these particularities exert their influence over timescales much shorter than the coalescent timescale; *cf.*, *e.g.*, Wakeley (2013). This is also the reason why the Kingman coalescent still appears as limiting genealogy of the weak seedbank model of Kaj *et al.* (2001) mentioned in the Introduction.

This robustness has turned the Kingman coalescent into an extremely useful tool in population genetics. In fact, it can be considered the standard null model for neutral populations. Its success is also based on the fact that it allows a simple derivation of many population genetic quantities of interest, such as a formula for the expected number of segregating sites

$$\mathbb{E}[S] = \frac{\theta}{2} \sum_{i=1}^{n-1} \frac{1}{i} =: \frac{\theta}{2} a(n) \tag{2}$$

or the expected average number of pairwise differences π (Tajima 1983), the expected values of the site-frequency spectrum (*cf.* Fu 1995), when one assumes the infinite-sites model of Watterson (1975). This analytic tractability has allowed the construction of a sophisticated statistical machinery for the inference of evolutionary parameters. We investigate the corresponding quantities for the seedbank coalescent below.

The seedbank coalescent with mutation

Similar to the Kingman coalescent, the seedbank coalescent, mathematically introduced in Blath *et al.* (2015b), describes the ancestral lines of a sample taken from a population with a seedbank component. Here, we distinguish whether an ancestral line belongs to an active or dormant individual for any given point backward in time. The main difference from the Kingman coalescent is that as long as an ancestral line corresponds to a dormant individual (in the seedbank), it cannot coalesce with other lines, since reproduction and thus finding a common ancestor are possible only for active individuals.

The dynamics are now easily described as follows: If there are currently n active and m dormant lineages at some point in the past, each "active pair" may coalesce with the same probability, after an exponential time with rate $\binom{n}{2}$, entirely similar to a classical Kingman coalescent with currently n lineages. However, each active line becomes dormant at a positive rate $c > 0$ (corresponding to an ancestor who emerged from the seedbank), and each dormant line resuscitates, at a rate cK, for some $K > 0$. The parameter K reflects the relative size of the seedbank compared to the active population and is explained below in terms of an explicit underlying population model. Since dormant lines are prevented from merging, they significantly delay the time to the most recent common ancestor. This mechanism is reminiscent of a structured coalescent with two islands

(Notohara 1990; Herbots 1997), where lineages may merge only if they are in the same colony. Of course, if one samples a seedbank coalescent backward in time, one need specify not only the sample size, but also the number of sampled individuals from the active population (say n) and from the dormant population (say m).

In this article, we also consider mutations along the ancestral lines. As in the Kingman case we place them along the active line segments according to a Poisson process with rate $\theta_1/2$ and along the dormant segments at a rate $\theta_2/2 \geq 0$. Depending on the concrete situation, one may want to choose $\theta_2 = 0$. To determine the mutation rate in dormant individuals will be an interesting inference question. In Figure 1, we illustrate a realization of the seedbank coalescent with mutations: Ancestral lineages residing in the seedbank are represented by dotted lines and do not take part in coalescences.

A formal mathematical definition of this process as a partition-valued Markov chain can be found in Blath *et al.* (2015b); it is straightforward to extend their framework to include mutations.

The parameters c and K can be understood as follows: c describes the proportion of individuals that enter the seedbank per (macroscopic) coalescent time unit. It is thus the rate at which individuals become dormant. If the ratio of the size of the active population and the dormant population in the underlying population is $K:1$ (that is, the active population is K times the size of the dormant population), and absolute (and thus also relative) population sizes are assumed to stay constant, then, for the relative amount of active and dormant individuals to stay balanced, the rate at which dormant individuals resuscitate and return to the active population is necessarily of the form cK (see also Figure 2). It is important to note that in this setup, the average coalescent time that an inactive individual stays dormant is of the order $N/(cK)$. We later also include a positive mortality rate for dormant individuals; this will lead to a reduced "effective" relative seedbank $\tilde{K}$.

Robustness and underlying assumptions of the seedbank coalescent

As for the Kingman coalescent, it is important to understand the underlying assumptions that make the seedbank coalescent a reasonable model for the genealogy of a population: Again, we assume the types in the population to be selectively neutral, so that there are no significant fitness differences. Further, we assume the population size N and the seedbank size M to be constant and to be of the same order; that is, there exists a $K > 0$ so that $N = K \cdot M$; *i.e.*, the ratio between active and dormant individuals is constant equal to $K : 1$. Finally, the rate at which an active individual becomes dormant should be c (on the macroscopic coalescent scale), so that necessarily the average time (in coalescent time units) that an individual stays dormant before being resuscitated becomes $1/(cK)$. If one includes a positive mortality rate in the seedbank, this will lead to a modified parameter $\tilde{K}$; see below.
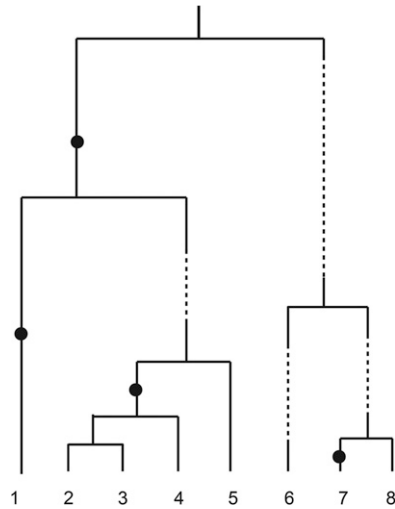
Figure 1 Realization of a seedbank coalescent with all $n = 8$ sampled lineages assumed active. Mutations are, in this example, allowed to occur only on active ancestral lineages represented by solid lines; ancestral lineages residing in the seedbank are represented by dotted lines. Coalescences are allowed only among active lineages.

We provide below an example of a concrete seedbank population model, the "Wright–Fisher model with geometric seedbank component," including mutation and mortality in the seedbank, for which it can be proved that the seedbank coalescent with mutation governs the genealogy if the population size N (and thus necessarily also seedbank size M) gets large, and coalescent time is measured in units of the population size N. This is the same scaling regime as in the case of the Kingman coalescent corresponding to genealogy of the classical Wright–Fisher model.

The seedbank coalescent with mutation should be robust against small alterations—such as in the transition or reproduction mechanism or in the population or seedbank size—of the underlying population, similar to the robustness of the Kingman coalescent, especially if these alterations occur on timescales that are much shorter than the coalescent timescale (which is N for the haploid Wright–Fisher model). For example, one can still obtain this coalescent in a *Moran model* with seedbank component, as long as the seedbank is on the same order as the active population and if the migration rates between seedbank and active population scale suitably (as well as the mutation rate) with the coalescent timescale. As mentioned above, this is an important difference from the model considered by Kaj *et al.* (2001), where the time an individual stays in the seedbank is negligible compared to the coalescent timescale, thus resulting merely in a (time change) of a Kingman coalescent—a weak seedbank effect.

A Wright–Fisher Model with Geometric Seedbank Distribution

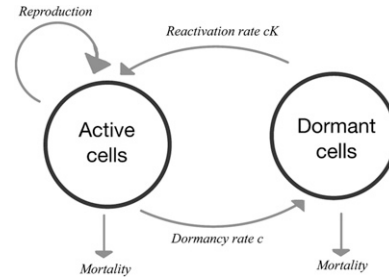We now introduce a Wright–Fisher-type population model with mutation and seedbank in which individuals

Figure 2 Dynamics of reversible microbial dormancy, according to Jones and Lennon (2010).

stay dormant for geometrically distributed amounts of time. The model is very much in line with classical probabilistic population genetics thinking (in particular, assuming constant population size), but also captures several features of microbial seedbanks described in Lennon and Jones (2011), in particular reversible states of dormancy and mortality in the seedbank. We assume that the following (idealized) aspects of (microbial) dormancy can be observed:

i. Dormancy generates a seedbank consisting of a reservoir of dormant individuals.

ii. The size of the seedbank is comparable to the order of the total population size, say in a constant ratio $K:1$ for some $K > 0$.

iii. The size of the active population N and of the seedbank $M = M(N)$ stays constant in time; combining this with ii we get $N = K \cdot M$.

iv. The model is selectively neutral so that reproduction is entirely symmetric for all individuals; for concreteness we assume reproduction according to the Wright–Fisher mechanism in fixed generations. That means the joint offspring distribution of the parents in each generation is symmetric multinomial. We interpret zero offspring as the death of the parent, one offspring as mere survival of the parent, and two or more offspring as successful reproduction leading to new individuals created by the parent.

v. Mutations may happen in the active population, at constant probability of the order $\theta_1/(2N)$, but potentially also in the dormant population [at the same, a reduced, or a vanishing probability $\theta_2/(2N)$].

vi. There is bidirectional and potentially repeated switching from active to dormant states, which appears essentially independently among individuals ("spontaneous switching"). The individual initiation probability of dormancy per generation is of the order c/N, for $c > 0$.

vii. Dormant individuals may die in the seedbank (due to maintenance and energy costs). If mortality is assumed to be positive, the individual probability of death per generation is of order d/N.

viii. For each new generation, all these mechanisms occur independently of the previous generations.

We schematically visualize this mechanism in Figure 2, which is similar to figure 1 in Jones and Lennon (2010). Whether these assumptions are met of course needs to be determined for the concrete underlying real population. In this theoretical article, we use the above assumptions to construct an explicit mathematical model that leads, measuring time in units of N, to a seedbank coalescent with mutation. Still, we emphasize that, as discussed in the previous section, the seedbank coalescent is robust as long as certain basic assumptions are met.

We now turn the above features into a formal mathematical model that can be rigorously analyzed, extending the Wright–Fisher model with a geometric seedbank component in Blath *et al.* (2015b) by additionally including mortality in the seedbank and potentially different mutation rates in the active and dormant populations.

Definition 1 (seedbank model with mutation and mortality)

Let $N \in \mathbb{N}$, and let $c, K, \theta_1 > 0$ and $\theta_2, d \geq 0$. The seedbank model with mutation is obtained by iterating the following dynamics for each discrete generation $k \in \mathbb{N}_0$ (with the convention that all occurring numbers are integers; if not, one may enforce this using appropriate Gauss brackets):

The N active individuals from generation $k = 0$ produce $N - c = N(1 - c/N)$ active individuals in generation $k = 1$ by multinomial sampling with equal weights.

Additionally, c dormant individuals, sampled uniformly at random without replacement from the seedbank of size $M := N/K$ in generation 0, reactivate; that is, they turn into exactly one active individual in generation $k = 1$ each and leave the seedbank.

The active individuals from generation 0 are thus replaced by these $(N - c) + c = N$ new active individuals, forming the active population in the next generation $k = 1$.

In the seedbank, d individuals, sampled uniformly at random without replacement from generation $k = 0$, die.

To replace the $c + d$ vacancies in the seedbank, the N active individuals from generation 0 produce $c + d$ seeds by multinomial sampling with equal weights, filling the vacant slots of the seeds that were activated.

The remaining $M - c - d = N/K - c - d$ seeds from generation 0 remain inactive and stay in the seedbank.

During reproduction, each newly created individual copies its genetic type from its parent.

In each generation, each active individual is affected by a mutation with probability $\theta_1/(2N)$, and each dormant individual mutates with probability $\theta_2/(2N)$ (where θ_2 may be 0).

This model is an extension of the model in Blath *et al.* (2015b) to additionally include mortality in the seedbank and incorporate (potentially distinct) mutation rates in the active and dormant population. It appears to be a rather natural extension of the classical Wright–Fisher model. Note

that the model has a geometric seedbank age distribution, since every dormant individual in each generation has the same probability to become active, resp. die, in the next generation, so that the time that an individual is in the dormant state is geometrically distributed. The parameter of this geometric distribution is given by

$$\frac{c}{M} = \frac{cK}{N} \qquad \text{resp.} \qquad \frac{c+d}{M} = \frac{(c+d)K}{N}$$

in the absence, resp. presence, of mortality in the seedbank. With mathematical arguments similar to those applied in Blath *et al.* (2015b), it is now standard to show that the ancestral process of a sample taken from the above population model converges, on the coalescent timescale N, to the seedbank coalescent with parameters c and K, resp.

$$\tilde{K} := \frac{c+d}{c} K,$$

and mutation rates $\theta_1/2, \theta_2/2$. It is interesting to see that mortality leads to a decrease of the relative seedbank size in a way that depends on the initiation rate c, which is of course rather intuitive. In this sense $\tilde{K}$ gives the effective relative seedbank size.

The type frequencies in the biallelic seedbank population model

In this article, we mostly consider the *infinite-sites model* (Watterson 1975), where it is assumed that each mutation generates an entirely new type. However, before turning to the infinite-sites model, we briefly discuss the biallelic case, say with types $\{a, A\}$. Given initial type configurations $\xi_0 \in \{a, A\}^N$ and $\eta_0 \in \{a, A\}^M$, denote by

$$\xi_k := (\xi_k(i))_{i \in [N]} \quad \text{and} \quad \eta_k := (\eta_k(j))_{j \in [M]}, \quad k \in \mathbb{N},$$

the genetic type configuration of the active individuals (ξ) and the dormant individuals (η) in generation k (obtained from the above mechanism). We assume that each mutation causes a transition from a to A or from A to a. Let

$$X_k^N := \frac{1}{N} \sum_{i \in [N]} 1_{\{\xi_k(i) = a\}} \quad \text{and} \quad Y_k^M := \frac{1}{M} \sum_{j \in [M]} 1_{\{\eta_k(j) = a\}},$$
$$k \in \mathbb{N}_0.$$

(3)

We call the discrete-time Markov chain $(X_k^N, Y_k^M)_{k \in \mathbb{N}_0}$ the *Wright–Fisher frequency process with mutation and seedbank component*. It can be seen from a generator computation that under our assumption with time measured in units of the active population size N it converges as $N \to \infty$ to the two-dimensional diffusion $(X_t, Y_t)_{t \geq 0}$ that is the solution to the system of stochastic differential equations

$$dX_t = \frac{\theta_1}{2}(1 - X_t)dt - \frac{\theta_1}{2}X_t dt + c(Y_t - X_t)dt$$
$$+ \sqrt{X_t(1 - X_t)}dB_t, \qquad (4)$$
$$dY_t = \frac{\theta_2}{2}(1 - Y_t)dt - \frac{\theta_2}{2}Y_t dt + (c + d)K(X_t - Y_t)dt.$$

Here, $(B_t)_{t \geq 0}$ denotes standard one-dimensional Brownian motion. An alternative way to represent this stochastic process is via its Kolmogorov backward generator (*cf.*, *e.g.*, Karlin and Taylor 1981), which is given by

$$\mathcal{L}f(x,y) = \frac{\partial f(x,y)}{\partial x}\left[\frac{\theta_1}{2}(1 - x) - \frac{\theta_1}{2}x + c(y - x)\right]$$
$$+ \frac{1}{2}\frac{\partial^2 f(x,y)}{\partial x^2}x(1 - x)$$
$$+ \frac{\partial f(x,y)}{\partial y}\left[\frac{\theta_2}{2}(1 - y) - \frac{\theta_2}{2}y + (c + d)K(x - y)\right]$$

for functions $f \in C^2([0,1]^2)$. Existence and uniqueness of the stationary distribution of this process follow from compactness of the state space and strictly positive mutation rate $\theta_1 > 0$ in the active population. See Chap. 4.9 in Ethier and Kurtz (2005) for more detailed arguments. Note that $\mathcal{L}$ is reminiscent of the backward generator of the *structured coalescent* with two islands (Notohara 1990; Herbots 1997); however, its qualitative behavior is very different. Its relation to the structured coalescent with two islands will be investigated in future research. Observe that the solution of (4) is driven by only one Brownian motion.

Population Genetics with the Seedbank Coalescent

In contrast to Lennon and Jones (2011), who use a deterministic population dynamics approach to study seedbanks, we are interested in probabilistic effects of seedbanks on genetic variability. We therefore use a coalescent approach to study the (random) gene genealogy of a sample. To better understand how seedbanks shape genealogies, we consider genealogical properties, such as the time to the most recent common ancestor, the total tree size, and the length of external branches.

Genealogical tree properties

First we discuss some classical population genetic properties of the seedbank coalescent when viewed as a random tree without mutations. For the results that we derive below, it will usually be sufficient to consider the *block-counting process* $(N_t, M_t)_{t \geq 0}$, of our coalescent, where N_t gives the number of lines in our coalescent that are active and M_t denotes the number of dormant lines t time units in the past. Then, $(N_t, M_t)_{t \geq 0}$ is the continuous-time Markov chain started in $(N_0, M_0) \in \mathbb{N}_0 \times \mathbb{N}_0$ with transitions

$$(n,m) \mapsto \begin{cases} (n-1, m+1), & \text{at rate } cn, \\ (n+1, m-1), & \text{at rate } (c + d)Km = \tilde{c}Km, \\ (n-1, m), & \text{at rate } \binom{n}{2}. \end{cases}$$
$$(5)$$

Again, introducing mutation can be done in the usual way, by superimposing independent Poisson processes with rate $\theta_1/2$ on the active lines and at rate $\theta_2/2$ on the dormant lines. If the block-counting process is currently in state $(N_t, M_t) = (n, m)$, then a mutation in an active line happens at rate $n\theta_1/2$ and a mutation in a dormant line at rate $m\theta_2/2$. The total jump rate from state (n, m) of the *backward process with mutation* is thus given by

$$r_{n,m} := \binom{n}{2} + cn + (c + d)Km + \frac{\theta_1}{2}n + \frac{\theta_2}{2}m. \qquad (6)$$

Time to the most recent common ancestor

It has been shown in Blath *et al.* (2015b, theorem 4.6) that the expected time to the most recent common ancestor $(\mathbb{E}_{n,0}[T_{\mathrm{MRCA}}])$ for the seedbank coalescent, if started in a sample of active individuals of size n, is $O(\log \log n)$, in stark contrast to the corresponding quantity for the classical Kingman coalescent, which is bounded by 2, uniformly in n; *cf.* (1). This already indicates that one should expect elevated levels of (old) genetic variability under the seedbank coalescent, since more (old) mutations can be accumulated. While the above result shows the asymptotic behavior of the $\mathbb{E}_{n,0}[T_{\mathrm{MRCA}}]$ for large n, it does not give precise information for the exact absolute value, in particular for "small to medium" n. Here, we provide recursions for its expected value and variance that can be computed efficiently. First, we introduce some notation.

We define the *time to the most recent common ancestor* of the seedbank coalescent formally to be

$$T_{\mathrm{MRCA}} := \inf\{t > 0 : N_t + M_t = 1\}.$$

If the sample consists of *an* active and *bn* dormant individuals, for some $a, b \in \mathbb{R}^+$, then the expected time to the most recent common ancestor is $\log(bn + \log an)$ (Blath *et al.* 2015b). Here, it is interesting to note that the time to the most recent common ancestor of the Bolhausen–Sznitman coalescent is also $O(\log \log n)$ (Goldschmidt and Martin 2005). The Bolthausen–Sznitman coalescent is often used as a model for selection; *cf.*, *e.g.*, Neher and Hallatschek (2013).

One can compute the expected time to most recent common ancestor recursively as follows. For $n, m \in \mathbb{N}_0$ let

$$t_{n,m} := \mathbb{E}_{n,m}[T_{\mathrm{MRCA}}], \qquad (7)$$

where $\mathbb{E}_{n,m}$ denotes expectation when started in $(N_0, M_0) = (n, m)$, *i.e.*, with n *active* lines and m *dormant* ones. Observe

that we need to consider both types of lines to calculate $t_{n,m}$. Write

$$\lambda_{n,m} := \binom{n}{2} + cn + (c+d)Km, \tag{8}$$

and abbreviate

$$\alpha_{n,m} := \frac{\binom{n}{2}}{\lambda_{n,m}}, \quad \beta_{n,m} := \frac{cn}{\lambda_{n,m}}, \quad \gamma_{n,m} := \frac{(c+d)Km}{\lambda_{n,m}}. \tag{9}$$

Then we have the recursive representation

$$\mathbb{E}_{n,m}[T_{MRCA}] = t_{n,m} = \lambda_{n,m}^{-1} + \alpha_{n,m}t_{n-1,m} + \beta_{n,m}t_{n-1,m+1} + \gamma_{n,m}t_{n+1,m-1}, \tag{10}$$

with initial conditions $t_{1,0} = t_{0,1} = 0$. The proof of (10) and a recursion for the variance of T_{MRCA} is given in Supporting Information, File S1. Since the process $N_t + M_t$ is nonincreasing in t, these recursions can be solved iteratively. In fact,

$$t_{2,0} = \frac{(c + (c+d)K)^2}{(c+d)^2 K^2}, \tag{11}$$

which in the case without mortality ($d = 0$) reduces to

$$t_{2,0} = 1 + \frac{2}{K} + \frac{1}{K^2}. \tag{12}$$

Notably, $t_{2,0}$ is constant for sample size 2 (see Equation 11) as c varies (Table 1) if $d = 0$ and in particular does not converge for $c \to 0$ to the Kingman case. This effect is similar to the corresponding behavior of the structured coalescent with two islands if the migration rate goes to 0; *cf.* Nath and Griffiths (1993). However, the Kingman coalescent values are recovered as the seedbank size decreases (*e.g.*, for $K = 100$ in Table 1).

The fact that $t_{2,0} = 4$ for $K = 1, d = 0$ can be understood heuristically if c is large: In that situation, transitions between active and dormant states happen very fast; thus at any given time the probability that a line is active is $\sim 1/2$, and therefore the probability that both lines of a given pair are active (and thus able to merge) is $\sim 1/4$. We can therefore conjecture that for $d = 0, K = 1$, and $c \to \infty$ the genealogy of a sample is given by a time change by a factor 4 of Kingman's coalescent.

Table 1 and Table 2 show values of $t_{n,0}$ obtained from (10) for various parameter choices and sample sizes. The relative size of the seedbank (K) has a significant effect on $\mathbb{E}_{n,0}[T_{MRCA}]$; a large seedbank ($K$ small) increases $\mathbb{E}_{n,0}[T_{MRCA}]$, while the effect of c is to dampen the increase in $\mathbb{E}_{n,0}[T_{MRCA}]$ with sample size (Table 1). The effect of the seedbank death rate d on $\mathbb{E}_{n,0}[T_{MRCA}]$ is to dampen the effect of the relative size (K) of the seedbank (Table 2).

Table 1 The expected time to the most recent common ancestor ($\mathbb{E}_{n,0}[T_{MRCA}]$) of the seedbank coalescent, obtained from (10), with seedbank size K, sample size n, dormancy initiation rates c as shown, and $d = 0$

	Sample size n		
c	2	10	100
$K = 0.01$, $\times 10^4$			
0.01	1.02	2.868	5.185
0.1	1.02	2.731	4.487
1	1.02	2.187	2.666
10	1.02	1.878	2.085
100	1.02	1.84	2.026
$K = 1$			
c	2	10	100
0.01	4	10.21	17.18
0.1	4	9.671	14.97
1	4	8.071	10.02
10	4	7.317	8.221
100	4	7.212	7.954
$K = 100$			
c	2	10	100
0.01	1.02	1.846	2.052
0.1	1.02	1.838	2.026
1	1.02	1.836	2.02
10	1.02	1.836	2.02
100	1.02	1.836	2.02
$K = \infty$	1	1.80	1.98

All sampled lines are from the active population [sample configuration $(n, 0)$]. For comparison, $\mathbb{E}_{(n)}[T_{MRCA}] = 2(1 - 1/n)$ when associated with the Kingman coalescent ($K = \infty$). The multiplication $\times 10^4$ applies only to the first section with $K = 0.01$.

Total tree length and length of external branches

To investigate the genetic variability of a sample, in terms, *e.g.*, of the number of segregating sites and the number of singletons, it is useful to have information about the total tree length and the total length of external branches. Let $L^{(a)}$ denote the total length of all branches while they are active and $L^{(d)}$ denote the total length of all branches while they are dormant. Their expectations

$$l_{n,m}^{(a)} := \mathbb{E}_{n,m}[L^{(a)}], \quad l_{n,m}^{(d)} := \mathbb{E}_{n,m}[L^{(d)}] \tag{13}$$

may be calculated using the following recursions for $n, m \in \mathbb{N}_0$, and with $\lambda_{n,m}$ given by (8),

$$l_{n,m}^{(a)} = n\lambda_{n,m}^{-1} + \alpha_{n,m}l_{n-1,m}^{(a)} + \beta_{n,m}l_{n-1,m+1}^{(a)} + \gamma_{n,m}l_{n+1,m-1}^{(a)}, \tag{14}$$

$$l_{n,m}^{(d)} = m\lambda_{n,m}^{-1} + \alpha_{n,m}l_{n-1,m}^{(d)} + \beta_{n,m}l_{n-1,m+1}^{(d)} + \gamma_{n,m}l_{n+1,m-1}^{(d)}. \tag{15}$$

Observe that Equations 14 and 15 differ in the factor (n resp. m) that multiplies $\lambda_{n,m}^{-1}$. Similar recursions hold for their variances as well as for the corresponding values of

Table 2 The expected time to most recent common ancestor ($\mathbb{E}_{n,0}[T_{\mathrm{MRCA}}]$) of the seed-bank coalescent, obtained from (10), with all $n = 100$ sampled lines assumed active and c, K, and d as shown

	$c = 1$, $n = 100$, parameter d				
K	0.01	0.1	1	10	100
0.01	2.614e+04	2.208e+04	6814	270.7	9.91
0.1	315.6	270.7	96.2	9.04	2.442
1	9.91	9.04	5.201	2.4	2.02
10	2.442	2.4	2.197	2.017	1.984
100	2.02	2.017	2	1.984	1.98
	$K = 1$, $n = 100$, parameter d				
c	0.01	0.1	1	10	100
0.01	8.281	2.893	2.051	1.985	1.98
0.1	13.39	7.215	2.617	2.025	1.984
1	9.91	9.04	5.201	2.4	2.02
10	8.213	8.138	7.477	4.556	2.361
100	7.953	7.946	7.875	7.245	4.466

For comparison, $\mathbb{E}_{(n)}[T_{\mathrm{MRCA}}] = 2(1 - 1/n)$ (1.98 for $n = 100$) when associated with the Kingman coalescent.

the total length of external branches, which can be found in File S2 together with the respective proofs. From (14) and (15) one readily obtains

$$l_{2,0}^{(a)} = \frac{2(c + (c + d)K)}{(c + d)K}, \quad l_{2,0}^{(d)} = \frac{2c(c + (c + d)K)}{(c + d)^2 K^2}. \tag{16}$$

We observe that $l_{2,0}^{(d)}$ and $l_{2,0}^{(a)}$ given in (16) are independent of c if $d = 0$ as also seen for $t_{2,0}$; *cf.* (11). We use (16) to obtain closed-form expressions for expected average number of pairwise differences.

The numerical solutions of (14) and (15) indicate that for $n \geq 2$,

$$l_{n,0}^{(a)} = \frac{(c + d)K}{c} \cdot l_{n,0}^{(d)} = \tilde{K} \cdot l_{n,0}^{(d)}. \tag{17}$$

Hence, the expected total lengths of the active and the dormant parts of the tree are proportional, and the ratio is given by the effective relative seedbank size.

Recursions for the expected total length of external branches are given in Proposition S1.3 in File S1. Let $e_{n,m}^{(a)}$ and $e_{n,m}^{(d)}$ denote the expected total lengths of active and dormant external branches, respectively, when started with n active and m dormant lines. The numerical solutions of the recursions indicate that the ratio of expected values $e_{n,0}^{(a)}$ and $e_{n,0}^{(d)}$ is also given by (17).

Recursions for expected branch lengths associated with any other class than singletons are more complicated to derive, and we postpone those for further study. Simulation results (not shown) suggest that the result (17) we obtained for relative expected total length of active branches, and active external branches, holds for all branch length classes; if $B_i^{(a)}$ ($B_i^{(d)}$) denotes the total length of active (dormant) branches subtending $i \in \{1, 2, \dots, n-1\}$ leaves, then, if all our sampled lines are active, we claim that

$\mathbb{E}[B_i^{(a)}]/(\mathbb{E}[B_i^{(a)}] + \mathbb{E}[B_i^{(d)}])$ is given by (17). Table S1 shows values of $r_{10,10} := e_{10,10}^{(a)}/(e_{10,10}^{(a)} + e_{10,10}^{(d)})$, *i.e.*, the relative expected total length of external branches when our sample consists of 10 active lines and 10 dormant ones. In contrast to the case when all sampled lines are active, c clearly affects $r_{10,10}$ when d is small. In line with previous results, d reduces the effect of the relative size (K) of the seedbank.

Table S2 shows the expected total lengths of active and dormant external branches $e_{n,0}^{(a)}$ and $e_{n,0}^{(d)}$ for values of c, K, and d as shown. When the seedbank is large (K small), $e_{n,0}^{(a)}$ and $e_{n,0}^{(d)}$ can be much longer than the expected length of 2 when associated with the Kingman coalescent (Fu 1995). However, as noted before, the effect of K depends on d. The effect of c also depends on d; changes in c have a bigger effect when d is large.

One can gain insight into the effects of a seedbank on the site-frequency spectrum by studying the effects of a seedbank on relative branch lengths. Let $R_i^{(a)} := B_i^{(a)}/B^{(a)}$ denote the relative total length of active branches subtending i leaves ($B_i^{(a)}$), relative to the total length of active branches $B^{(a)} = B_1^{(a)} + \cdots + B_{n-1}^{(a)}$, and we consider only the case when all n sampled lines are active. Thus, if one assumes that the mutation rate in the seedbank is negligible compared to the mutation rate in the active population, $\mathbb{E}_{n,0}[R_1^{(a)}]$ should be a good indicator of the relative number of singletons, relative to the total number of segregating sites. In addition, we investigate $\mathbb{E}_{n,0}[R_i^{(a)}]$ to learn whether and how the presence of a seedbank affects genetic variation, even if *no* mutations occur in the seedbank. Figure S1 shows estimates of $\mathbb{E}_{n,0}[R_i^{(a)}]$ (obtained by simulations) for values of c, K, and d as shown (all $n = 100$ sampled lines are assumed active). The main conclusion is that a large seedbank reduces the relative length of external branches and increases the relative magnitude of the right tail of the branch length spectrum. Thus, one would expect to see a similar pattern in neutral genetic variation: a reduced relative number of singletons (relative to the total number of polymorphic sites) and a relative increase in the number of polymorphic sites in high count.

Neutral genetic variation

In this subsection we derive and study several recursions for common measures of DNA sequence variation in the infinite-sites model (ISM) of Watterson (1975). Samples are assumed to be drawn from the stationary distribution. We also investigate how these quantities differ from the corresponding values under the Kingman coalescent, in an effort to understand how seedbank parameters affect genetic variability.

Segregating sites: First we consider the *number of segregating sites S in a sample*, which, assuming the ISM, is the total number of mutations that occur in the genealogy of the sample until the time of its most recent common ancestor. In addition to being of interest on its own, S is

a key ingredient in commonly employed distance statistics such as those of Tajima (1989) and Fu and Li (1993). We let mutations occur on active branch lengths according to independent Poisson processes each with rate $\theta_1/2$ and on dormant branches with rate $\theta_2/2$. The expected value of S can be expressed in terms of the expected total tree lengths as

$$\mathbb{E}_{n,m}[S] = \frac{\theta_1}{2} l_{n,m}^{(a)} + \frac{\theta_2}{2} l_{n,m}^{(d)}. \tag{18}$$

An alternative recursion for the expectation and the variance of the number of segregating sites can be found in File S1 (Proposition S1.5).

Table 3 shows the expected number of segregating sites $\mathbb{E}_{n,0}[S] = s_{n,0}$ in a sample of size n taken from the active population for values of c and K as shown. The size of the seedbank K strongly influences the number of segregating sites. If there is no mutation in the seedbank, it roughly doubles for $K = 1$ and approaches the normal value of the Kingman coalescent for small seedbanks ($K = 100$). The parameter K seems to have a more significant influence than the parameter c.

Average pairwise differences: Average pairwise differences are a key ingredient in the distance statistics of Tajima (1983) and Fay and Wu (2000). Expected value and variance for average pairwise differences in the Kingman coalescent were first derived in Tajima (1983). Here, we give an expression for the expectation in terms of the expected total tree lengths. Denote by π the average number of pairwise differences

$$\pi = \frac{1}{\binom{N_0 + M_0}{2}} K, \tag{19}$$

where $K = \sum_{(i,j):i<j} K_{ij}$ is the total number of pairwise differences, with K_{ij} denoting the number of differences observed in the pair of DNA sequences indexed by (i,j). We abbreviate $d_{n,m} := \mathbb{E}_{n,m}[K]$ and obtain

$$\mathbb{E}_{n,m}[\pi] = \frac{1}{\binom{n+m}{2}} d_{n,m},$$

which can be calculated using

$$\mathbb{E}_{n,m}[\pi] = \frac{1}{\binom{n+m}{2}} \left[\binom{n}{2} \left(\frac{\theta_1}{2} l_{2,0}^{(a)} + \frac{\theta_2}{2} l_{2,0}^{(d)} \right) \right.$$
$$\left. + nm \left(\frac{\theta_1}{2} l_{1,1}^{(a)} + \frac{\theta_2}{2} l_{1,1}^{(d)} \right) + \binom{m}{2} \left(\frac{\theta_1}{2} l_{0,2}^{(a)} + \frac{\theta_2}{2} l_{0,2}^{(d)} \right) \right], \tag{20}$$

where $l_{n,m}^{(a)}$ and $l_{n,m}^{(d)}$ are defined in (13).

Table 3 The expected total number of segregating sites ($s_{n,0}$), with values of K, c, and d as shown and sample size $n = 100$ (all lines from the active population), with $\theta_1 = 2$ and $\theta_2 = 0$

	Values of K, $d = \theta_2 = 0$				
c	0.01	0.1	1	10	100
0.01	1035	112.8	20.6	11.38	10.46
0.1	958.3	105.3	19.98	11.36	10.46
1	790.6	90.37	19.4	11.38	10.46
10	884	99.92	20.16	11.39	10.46
100	1010	110.8	20.61	11.39	10.46
	Values of K, $d = 100$, $\theta_2 = 0$				
c	0.01	0.1	1	10	100
0.01	10.46	10.37	10.36	10.35	10.35
0.1	11.36	10.46	10.37	10.36	10.35
1	19.32	11.37	10.46	10.37	10.36
10	91.97	19.3	11.29	10.45	10.36
100	510.3	60.83	15.51	10.87	10.41

When associated with the Kingman coalescent, $\theta = 2$ and $s_{(100)} = 10.35$.

Hence, given a sample configuration $(n, 0)$, *i.e.*, our n sampled lines are all active, (20), together with (16), gives

$$\mathbb{E}_{n,0}[\pi] = \frac{c + (c+d)K}{(c+d)K} \left(\theta_1 + \frac{c\theta_2}{(c+d)K} \right). \tag{21}$$

If now $d = 0$, the dependence on c disappears again, since we have

$$\mathbb{E}_{n,0}[\pi] = \theta_1 + \frac{\theta_1}{K} + \left(1 + \frac{1}{K} \right) \frac{\theta_2}{K},$$

which is obviously highly elevated compared to θ_1 if the seedbank is large (K small). For comparison, $\mathbb{E}_{(n)}[\pi] = \theta_1$ when associated with the usual Kingman coalescent, which we recover in the absence of a seedbank ($K \to \infty$) in (21).

The site-frequency spectrum: The site-frequency spectrum (SFS) is one of the most important summary statistics of population genetic data in the infinite-sites model. Suppose that we can distinguish between mutant and wild type, *e.g.*, with the help of an outgroup. As before, we distinguish between the number of samples taken from the active population (say n) and the dormant population (say m). Then, the SFS of an (n, m) sample is given by

$$\underline{\xi}^{(n,m)} := \left(\xi_1^{(n+m)}, \ldots, \xi_{n+m-1}^{(n+m)} \right), \tag{22}$$

where the $\xi_i^{(n+m)}$, $i = 1, \ldots, n+m-1$ denote the number of sites at which variants appear i times in our sample of size $n + m$. For the Kingman coalescent, the expected values, variances, and covariances of the SFS have been derived by Fu (1995). Expected values and covariances can in principle be computed by extending the theory in Fu (1995) and resp. Griffiths and Tavaré (1998). However, the computations would be far more involved than the previous

recursions and will be treated in future research. We derive recursions for the expected number of singletons and investigate the whole SFS by simulation.

Number of singletons: The number of singletons in a sample is often taken as an indicator of the kind of historical processes that have acted on the population. By "singletons" we mean the number of *derived* (or new) mutations that appear only once in the sample, which in the infinite-sites model is equal to the number of mutations occurring on external branches. Thus we can relate the expected number of singletons, denoted by $\xi_1^{(n+m)}$, to the total length of external branches in the same way as we related the number of segregating sites to the total tree length. Let $e_{n,n',m,m'}^{(a)}$ denote the expected total length of external branches when our sample consists of n *active external* lines, n' *active internal* lines, m *dormant external* lines, and m' *dormant internal* lines. Define $e_{n,n',m,m'}^{(d)}$ similarly as the expected total length of dormant external branches. Recursions for $e_{n,n',m,m'}^{(d)}$ and $e_{n,n',m,m'}^{(a)}$ are given in File S2. For $n, m \in \mathbb{N}_0$ we have that the expected number of singletons is given by

$$\xi_1^{(n+m)} = \frac{\theta_1}{2} e_{n,0,m,0}^{(a)} + \frac{\theta_2}{2} e_{n,0,m,0}^{(d)}.$$

Thus, one can compute the expected number of singletons by solving the recursions for external branch lengths. By way of example, Table S2 gives values of $e_{n,0,m,0}^{(a)}$ and $e_{n,0,m,0}^{(d)}$ for a sample of 10 active lines ($n = 10$, $m = 0$).

The whole site-frequency spectrum: Figure 3 shows estimates of the normalized expected frequency spectrum $\mathbb{E}\left[\xi_i^{(n,0)}\right] \Big/ \mathbb{E}\left[\left|\xi^{(n,0)}\right|\right]$, where $\left|\xi^{(n,0)}\right| = \xi_1^{(n,0)} + \cdots + \xi_{n-1}^{(n,0)}$ denotes the total number of segregating sites. Figure 3 shows that if the relative size of the seedbank is small (say, $K = 100$), then the SFS is almost unaffected by dormancy, in line with intuition. If the seedbank is large (say $K = 0.1$) and the transition rate $c = 1$ is comparable to the mutation rate $\theta_1/2 = 1$, then the spectrum differs significantly; in particular, the number of singletons is reduced by about one-half, which should be significant, and the right tail is much heavier.

This can be understood as follows: If the seedbank leads to an extended time to the most recent common ancestor, then the proportion of old mutations should increase, and these should be visible in many sampled individuals, strengthening the right tail of the spectrum.

It is interesting to see that even in the presence of a large seedbank (say $K = 0.1$), large transition rates (say $c = 100$) do not seem to affect the normalized spectrum. Again, this can be understood intuitively, since by the arguments presented in the discussion after (12) large c should lead to a constant time change of the Kingman coalescent (with a time change depending on K). Such a time change does not affect the normalized spectrum.

One reason for considering the SFS is naturally that one wants to be able to use the SFS in inference, to determine, say, if a seedbank is present and how large it is. If one has expressions for the expected SFS under some coalescent model, one can use the normalized expected SFS in an approximate-likelihood inference (see, *e.g.*, Eldon *et al.* 2015). The normalized spectrum is also appealing since it is quite robust to changes in the mutation rate (Eldon *et al.* 2015). For comparison, Figure S2 shows estimates of the expected normalized spectrum $\mathbb{E}[\zeta_i^{(n)}]$ where $\zeta_i^{(n)} := \xi_i^{(n)} / |\xi^{(n)}|$ and shows a similar pattern as for the normalized expected spectrum in Figure 3.

Distance statistics

Rigorous inference work is beyond the scope of this article. However, we can still consider (by simulation) estimates of the distribution of various commonly employed distance statistics. Distance statistics for the site-frequency spectrum are often employed to make inference about historical processes acting on genetic variation in natural populations. Commonly used statistics include the ones of Tajima (1989) (D_T), Fu and Li (1993) (D_{FL}), and Fay and Wu (2000) (D_{FW}). These statistics contrast different parts of the site-frequency spectrum (*cf.*, *e.g.*, Zeng *et al.* 2006).

The ℓ_2 distance: Arguably the most natural distance statistic to consider is the ℓ_2 distance (or sum of squares) of the whole SFS (or some lumped version thereof) between the observed SFS and an expected SFS based on some coalescent model. The $\ell_2^{(n)}$ statistic (n denotes sample size) is given by

$$\ell_2^{(n)} = \left(\sum_{i=1}^{n-1} \frac{\left(\xi_i^{(n)} - \mathbb{E}\left[\xi_i^{(n)}\right] \right)^2}{\mathrm{Var}\left[\xi_i^{(n)}\right]} \right)^{1/2}, \qquad (23)$$

where, in our case, expectation and variance are taken with respect to the Kingman coalescent (Fu 1995). Estimates of the distribution of $\ell_2^{(n)}$ are shown in Figure 4. As the size of the seedbank increases (K decreases), one observes worse fit of the site-frequency spectrum with the expected SFS associated with the Kingman coalescent.

Tajima's D: Tajima's statistic (D_T) for a sample of size n, with $a(n) = \sum_{j=1}^{n-1} j^{-1}$, is defined as

$$D_T = \frac{\pi - S/a(n)}{\sqrt{\mathbb{V}[\pi - S/a(n)]}} \qquad (24)$$

(Tajima 1989), where the variance $\mathbb{V}[\pi - S/a(n)]$ depends on the mutation rate θ that is usually estimated from the data. Under the Kingman coalescent, $\mathbb{E}[D_T] = 0$. Deviations from the Kingman coalescent model become significant at the 5% level if they are either >2 or < -2. Negative values of D_T should appear if there is an excess of either low- or high-frequency polymorphisms and a deficiency of
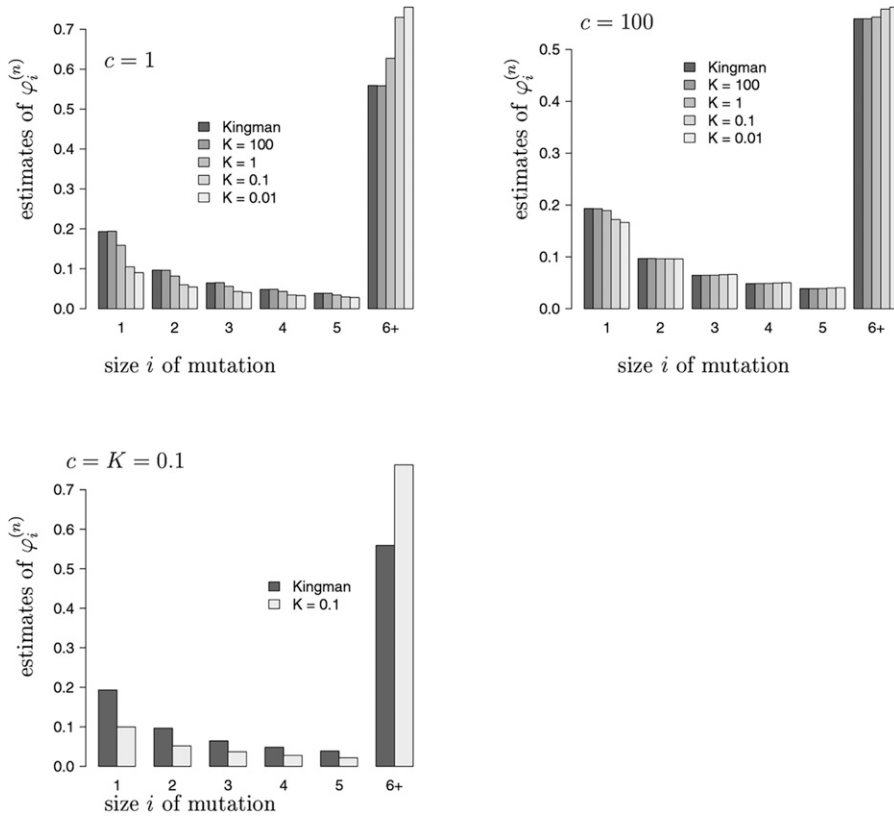
Figure 3 Estimates of the normalized site-frequency spectrum $\varphi_i^{(n)} = \mathbb{E}[\xi_i^{(n)}]/\mathbb{E}[|\xi^{(n)}|]$ with all $n = 100$ sampled lines assumed active and values of c and K as shown ($d = 0$). The mutation rate in the active population is fixed, $\theta_1 = 2$, and there is no mutation in the dormant states ($\theta_2 = 0$). All estimates based on 10^5 replicates.

middle-frequency polymorphisms (see, *e.g.*, Wakeley 2009 for further details). Positive values of D_T are to be expected if variation is common with moderate frequencies, for example in the presence of a recent population bottleneck or balancing selection.

The empirical distribution of D_T was investigated by simulation for different seedbank parameters (Figure 5 and Figure S3), assuming that mutations do not occur in the seedbank ($\theta_2 = 0$). If the seedbank is large ($K = 1/10$, $1/100$), then the median of D becomes significantly positive. For $c = K = 1$, there is very little deviation from the Kingman coalescent. Again D seems to be more sensitive to small values of K than changes in c. This is in line with our results on the $\mathbb{E}_{n,0}[T_{\mathrm{MRCA}}]$, with highly elevated times for small K. In the latter case, old variation will dominate, thus resembling a population bottleneck, producing positive values of D_T.

In conclusion, D_T might not be a very good statistic to detect seedbanks.

Fu and Li's D: Fu and Li (1993) statistic D_{FL} is defined as

$$D_{\mathrm{FL}} = \frac{S - a(n)\xi_1}{\sqrt{u_n S + v_n S^2}} \qquad (25)$$

with S being the total number of segregating sites, ξ_1 the total number of singletons, $a(n) = \sum_{j=1}^{n-1} j^{-1}$, and u_n and v_n as in Fu and Li (1993). As with D_T, $\mathbb{E}[D_{\mathrm{FL}}] = 0$ under the Kingman coalescent.

Figure 6 shows estimates of the distribution of D_{FL} assuming $\theta_2 = 0$. When the seedbank is large (K small), the distribution of D_{FL} becomes highly skewed, with most genealogies resulting in a low number of singletons compared with the total number of polymorphisms, resulting in positive D_{FL}. This is in line with our observations about the relative number of singletons associated with a large seedbank (Figure 3 and Figure S2) and the relative length of external branches (Figure S1).

Fay and Wu's H: The distance statistics D_{FW} of Fay and Wu (2000) are defined as

$$D_{\mathrm{FW}} = \frac{H - \pi}{\sqrt{\mathrm{Var}(H - \pi)}}, \qquad (26)$$

where

$$H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \xi_i^{(n)} i^2 \qquad (27)$$

and π is the average number of pairwise differences. A formula for the variance of D_{FW} was obtained by Zeng *et al.* (2006). Figure 7 holds estimates of the distribution of D_{FW} with $n = 100$, $d = 0$, and c and K as shown. As the seedbank size increases (K decreases), high-frequency variants, as captured by H, become dominant over the middle-frequency variants captured by π. In conclusion, Fu and Li's D_{FL} or Fay and Wu's D_{FW} may be preferable over Tajima's statistic D_T to
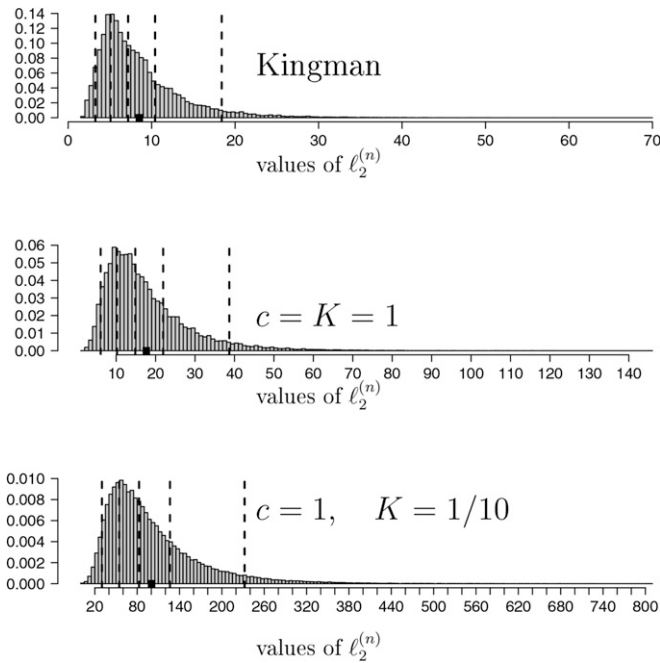
Figure 4 Estimates of the distribution of the $\ell_2^{(n)}$ statistic (Equation 23), with all $n = 100$ sampled lines assumed active, c and K as shown, $\theta_1 = 2$, and $\theta_2 = 0$. The vertical dashed lines are the 5%, 25%, 50%, 75%, and 95% quantiles and the solid square (■) denotes the mean. The entries are normalized to have unit mass 1. All estimates are based on 10^5 replicates.

detect the presence of a seedbank. A rigorous comparison of different statistics (including the E statistic of Zeng *et al.* 2006) and their power to distinguish between absence and presence of a seedbank must be the subject of future research.

The C code written for the computations is available at http://page.math.tu-berlin.de/~eldon/programs.html.

Discussion

In the previous sections, we have presented and analyzed an idealized model of a population sustaining a large seedbank, as well as the resulting patterns of genetic variability, with the help of a new coalescent structure, called the seedbank coalescent (with mutation). This ancestral process appeared naturally as a scaling limit of the genealogy of large populations producing dormant forms, in a similar way to the classical Kingman coalescent that arises in conventional models, under the following assumptions: The seedbank size is of the same order as the size of the active population, the population and seedbank size are constant over time, and individuals enter the dormant state by spontaneous switching independently of each other, in a way that individual dormancy times are comparable to the active population size. We begin with a discussion of these modeling assumptions.

The assumption that the seedbank is of comparable size to the active population is based on Lennon and Jones (2011), where it is shown in box 1, table a, that this is often the case in microbial populations.
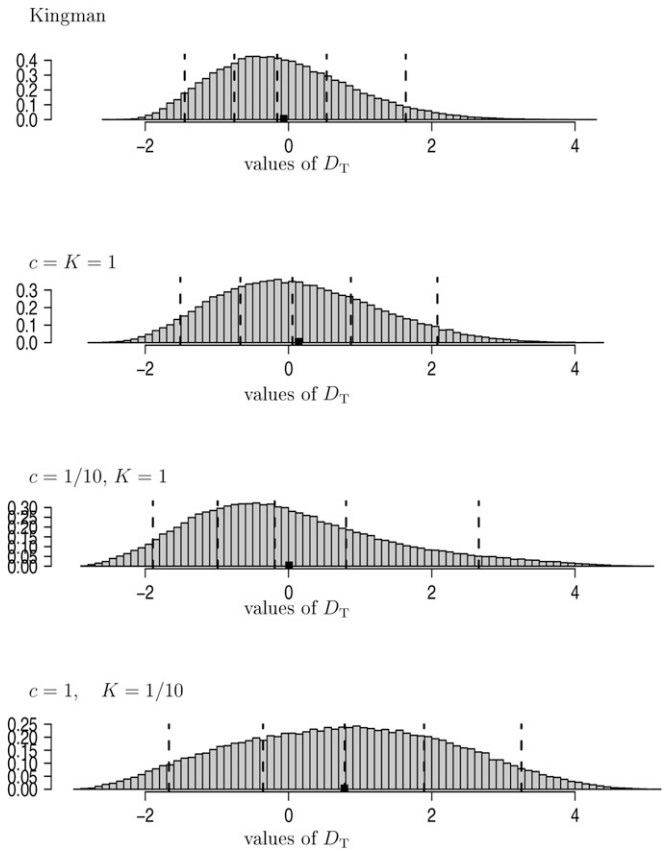

Figure 5 Estimates of the distribution of Tajima's D_T (Equation 24) with all $n = 100$ sampled lines assumed active, $\theta_1 = 2$, and $\theta_2 = 0$. The vertical dashed lines are the 5%, 25%, 50%, 75%, and 95% quantiles and the solid square (■) denotes the mean. The entries are normalized to have unit mass 1. The histograms are drawn on the same horizontal scale. Estimates are based on 10^5 replicates.

Assuming constant population size is a very common simplification in population genetics and can be explained with constant environmental conditions, we claim that "weak" fluctuations (of smaller order than the active population size) still lead to the seedbank coalescent model, as is the case for the Kingman coalescent. However, seedbanks are often seen as a bet-hedging strategy against drastic environmental changes, which is not yet covered by our models. We see this as an important task for future research, which will require serious mathematical analysis. In the case of weak seedbank effects, fluctuating population size has been considered in Živković and Tellier (2012), where the presence of the seedbank was observed to lead to an increase of the effective population size.

Assuming spontaneous switching of single individuals between the active and the dormant state is also based on Lennon and Jones (2011) (pp. 122–124). This somewhat restricts the scope of the model because it will not capture major environmental changes that may trigger a simultaneous change of state of a large proportion of individuals (*e.g.*, due to sudden lack of nutrients). This effect is closely related to drastic changes in population size and again may lead to serious alterations of our predictions. Hence,
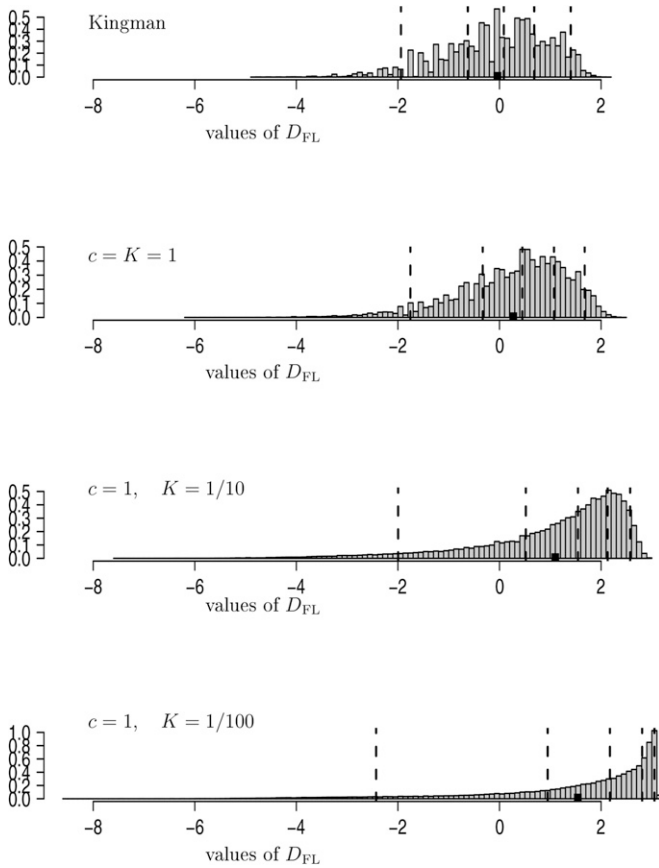
Figure 6 Estimates of the distribution of Fu and Li's D_{FL} (Equation 25) with all $n = 100$ sampled lines assumed active, $\theta_1 = 2$, and $\theta_2 = 0$. The vertical dashed lines are the 5%, 25%, 50%, 75%, and 95% quantiles and the solid square (■) denotes the mean. The entries are normalized to have unit mass 1. The histograms are drawn on the same horizontal scale. Estimates are based on 10^5 replicates.

Figure 7 Estimates of the distribution of Fay and Wu's D_{FW} (Equation 26) with all $n = 100$ sampled lines assumed active, $\theta_1 = 2$, and $\theta_2 = 0$. The vertical dashed lines are the 5%, 25%, 50%, 75%, and 95% quantiles and the solid square (■) denotes the mean. The entries are normalized to have unit mass 1. The histograms are drawn on the same horizontal scale. Estimates are based on 10^5 replicates.

including such large switching events will also be an important part of our future work (and will again require substantial mathematical work). In Vitalis *et al.* (2004) a whole proportion of the dormant population becomes active in every generation, but this should be seen in conjunction with the fact that dormancy is of limited duration, which excludes drastic alterations on a long timescale.

Assuming that the time spent in the seedbank is of the order of the population size is one of the main features that distinguishes our model from previous models of weak seedbank effects as previously investigated in Kaj *et al.* (2001) and Vitalis *et al.* (2004). Statistical inference will be needed to support or reject this assumption and to distinguish between weak and strong seedbanks. One distinguishing feature of weak and strong seedbanks is the behavior of the normalized site-frequency spectrum. Since weak seedbank effects lead to a genealogy that is a constant time change of Kingman's coalescent (Kaj *et al.* 2001; Blath *et al.* 2013), the normalized frequency spectrum of weak seedbanks will be similar to those corresponding to the Kingman coalescent, while under our model we observe (at least for large seedbanks) a reduction in the number of singletons (Figure S2).
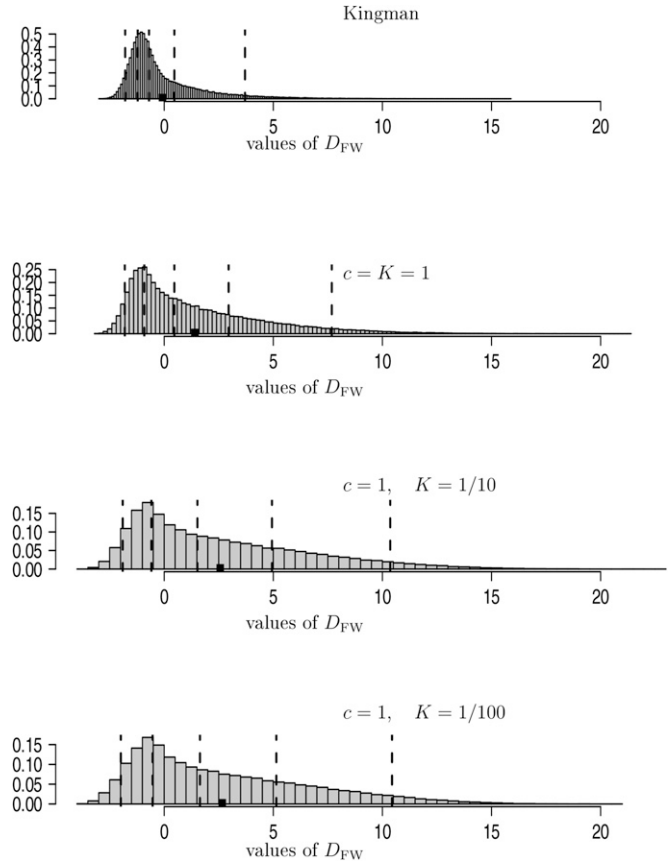
The model of Kaj *et al.* (2001) was used in Tellier *et al.* (2011), where Tajima's D was used to detect seedbanks.

We now discuss our results for the behavior of classical quantities describing genetic variability under our modeling assumptions, that is, when the genealogy of a sample can be described by the seedbank coalescent. In particular, we used it to derive recursions for quantities such as the time to the most recent common ancestor, the total tree length, or the length of external branches. We investigated statistics of interest to genetic variability such as the number of segregating sites, the site-frequency spectrum, Tajima's D, Fu and Li's D, and Fay and Wu's H by numerical solution of our recursions and by simulation. It turns out that the seedbank size K leads to significant changes, for example in the site frequency spectrum, producing a positive Tajima's D, indicating the presence of old genetic variability, in line with intuition. Interestingly, the influence of c seems to be less pronounced. For $K \to \infty$ we observe convergence toward the Kingman coalescent regime, while $c \to \infty$ seems to lead to a constant time change of Kingman's coalescent.

We are confident that our results so far have the potential to open up many interesting research questions, both on

modeling and on statistical inference, as well as in data analysis. For example, it should be interesting to derive a test to distinguish between the presence of strong *vs.* weak (resp. negligible) seedbanks. Another important task in future research will be to infer parameters of the model. While the relative seedbank size K can in principle be directly observed by cell counting (Lennon and Jones 2011), the parameter c seems to be difficult to observe, in particular because we have seen that many statistics we calculated are independent of or at least not very sensitive with respect to c. On the other hand, this shows that our results are fairly robust under alterations of c, such that estimations or tests may be applied to some extent without prior knowledge on c. The mortality rate d may for many practical purposes be included into the parameter K or $\tilde{K}$ measuring the "effective" relative seedbank size.

Estimating the mutation rates θ_1 and θ_2 is another goal for the future. In particular, in view of an ongoing debate on the possibility of mutations in dormant individuals (Maughan 2007), it would be important to devise a test to determine whether $\theta_2 > 0$.

Acknowledgments

Literature Cited

Blath, J., A. González Casanova, N. Kurt, and D. Spanò, 2013 The ancestral process of long-range seed bank models. J. Appl. Probab. 50: 741–759.

Blath, J., B. Eldon, A. Casanova, and N. Kurt, 2015a Genealogy of a Wright Fisher model with strong seed bank component, in *XI Symposium on Probability and Stochastic Processes, Progress in Probability*, edited by R. Mena, J. C. Pardo, V. Rivero, and G. Uribe Bravo. Birkhäuser, Basel, Switzerland (in press).

Blath, J., A. Gonzalez-Casanova, N. Kurt, and M. Wilke-Berenguer, 2015b The seed-bank coalescent. Ann. Appl. Probab. (in press).

Cannings, C., 1974 The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. Adv. Appl. Probab. 6: 260–290.

Eldon, B., M. Birkner, J. Blath, and F. Freund, 2015 Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? Genetics 199: 841–856.

Ethier, S. N., and T. G. Kurtz, 2005 *Markov Processes: Characterization and Convergence*. Wiley, Hoboken, NJ.

Fay, J. C., and C. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford.

Fu, Y.-X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. 48: 172–197.

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

Goldschmidt, C., and J. B. Martin, 2005 Random recursive trees and the Bolthausen-Sznitman coalescent. Electron. J. Probab. 10: 718–745.

Griffiths, R. C., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Special issue in honor of Marcel F. Neuts. Comm. Stat. Stoch. Models 14: 273–295.

Herbots, H. M., 1997 The structured coalescent, pp. 231–255 in *Progress of Population Genetics and Human Evolution*, edited by P. Donnelly and S. Tavaré. Springer-Verlag, Berlin/Heidelberg, Germany/New York.

Jones, S. E., and J. T. Lennon, 2010 Dormancy contributes to the maintenance of microbial diversity. Proc. Natl. Acad. Sci. USA 107: 5881–5886.

Kaj, I., S. M. Krone, and M. Lascoux, 2001 Coalescent theory for seed bank models. J. Appl. Probab. 38: 285–300.

Karlin, S., and H. Taylor, 1981 *A Second Course in Stochastic Processes*. Academic Press, New York.

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

Kingman, J. F. C., 1982b Exchangeability and the evolution of large populations, pp. 97–112 in *Exchangeability in Probability and Statistics*, edited by G. Koch, and F. Spizzichino. North-Holland, Amsterdam.

Kingman, J. F. C., 1982c On the genealogy of large populations. J. Appl. Probab. 19A: 27–43.

Lennon, J. T., and S. E. Jones, 2011 Microbial seed banks: the ecological and evolutionary implications of dormancy. Nat. Rev. Microbiol. 9: 119–130.

Maughan, H., 2007 Rates of molecular evolution in bacteria are relatively constant despite spore dormancy. Evolution 61: 280–288.

Moran, P., 1958 Random processes in genetics. Proc. Camb. Philos. Soc. 54: 60–71.

Nath, H., and R. Griffiths, 1993 The coalescent in two colonies with symmetric migration. J. Math. Biol. 31: 841–852.

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

Notohara, M., 1990 The coalescent and the genealogical process in geographically structured population. J. Math. Biol. 29: 59–75.

Tajima, F., 1983 Evolutionary relationships of DNA sequences in finite populations. Genetics 105: 437–460.

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

Tellier, A., S. J. Laurent, H. Lainer, P. Pavlidis, and W. Stephan, 2011 Inference of seed bank parameters in two wild tomato species using ecological and genetic data. Proc. Natl. Acad. Sci. USA 108: 17052–17057.

Vitalis, R., S. Glémin, and I. Olivieri, 2004 When genes got to sleep: the population genetic consequences of seed dormancy and monocarpic perenniality. Am. Nat. 163: 295–311.

Wakeley, J., 2009 *Coalescent Theory: An introduction*, Vol. 1. Roberts & Company Publishers, Greenwood Village, CO.

Wakeley, J., 2013 Coalescent theory has many new branches. Theor. Popul. Biol. 87: 1–4.

Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276.

Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97–159.

Zeng, K., Y. Fu, S. Shi, and C. Wu, 2006 Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174: 1431–1439.

Živković, D., and A. Tellier, 2012 Germ banks affect the inference of past demographic events. Mol. Ecol. 21: 5434–5446.

Communicating editor: W. Stephan

GENETICS

Genetic Variability Under the Seedbank Coalescent

Jochen Blath, Adrián González Casanova, Bjarki Eldon, Noemi Kurt, and Maite Wilke-Berenguer

File S1   Proofs and further recursive formulas

Expectation and variance of the T_{MRCA}

For $n, m \in \mathbb{N}_0$, let $t_{n,m} := \mathbb{E}_{n,m}[T_{\mathrm{MRCA}}]$ and $v_{n,m} := \mathbb{V}_{n,m}[T_{\mathrm{MRCA}}]$.

Proposition S1.1. *Let $n, m \in \mathbb{N}_0$. Then we have the following recursive representations*

$$\mathbb{E}_{n,m}[T_{\mathrm{MRCA}}] = t_{n,m} = \lambda_{n,m}^{-1} + \alpha_{n,m} t_{n-1,m} + \beta_{n,m} t_{n-1,m+1} + \gamma_{n,m} t_{n+1,m-1}, \tag{S1}$$

$$\begin{aligned}
\mathbb{V}_{n,m}[T_{\mathrm{MRCA}}] = v_{n,m} = {} & \lambda_{n,m}^{-2} + \alpha_{n,m} v_{n-1,m} + \beta_{n,m} v_{n-1,m+1} + \gamma_{n,m} v_{n+1,m-1} \\
& + \alpha_{n,m} t_{n-1,m}^2 + \beta_{n,m} t_{n-1,m+1}^2 + \gamma_{n,m} t_{n+1,m-1}^2 \\
& - \left(\alpha_{n,m} t_{n-1,m} + \beta_{n,m} t_{n-1,m+1} + \gamma_{n,m} t_{n+1,m-1} \right)^2,
\end{aligned} \tag{S2}$$

with initial conditions $t_{1,0} = t_{0,1} = v_{1,0} = v_{0,1} = 0$.

Proof of Proposition S1.1. Let τ_1 denote the time of the first jump of the process $(N_t, M_t)_{t \geq 0}$. If started at (n, m), this is an exponential random variable with parameter $\lambda_{n,m}$. Applying the strong Markov property we obtain

$$\begin{aligned}
t_{n,m} = {} & \mathbb{E}_{n,m}[\tau_1] + \mathbb{E}_{n,m}\left[\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[T_{\mathrm{MRCA}}] \right] \\
= {} & \lambda_{n,m}^{-1} + \alpha_{n,m} t_{n-1,m} + \beta_{n,m} t_{n-1,m+1} + \gamma_{n,m} t_{n+1,m-1}.
\end{aligned}$$

Similarly, the strong Markov property (telling us that τ_1 is independent of the time to the most recent common ancestor of the (random) sample (N_{τ_1}, M_{τ_1})) and the law of total

variance yields

$$v_{n,m} = \mathbb{V}_{n,m}[\tau_1] + \mathbb{E}_{n,m}\big[\mathbb{V}_{N_{\tau_1},M_{\tau_1}}[T_{\mathrm{MRCA}}]\big] + \mathbb{V}_{n,m}\big[\mathbb{E}_{N_{\tau_1},M_{\tau_1}}[T_{\mathrm{MRCA}}]\big]$$
$$= \lambda_{n,m}^{-2} + \mathbb{E}_{n,m}\big[\mathbb{V}_{N_{\tau_1},M_{\tau_1}}[T_{\mathrm{MRCA}}]\big] + \mathbb{V}_{n,m}\big[\mathbb{E}_{N_{\tau_1},M_{\tau_1}}[T_{\mathrm{MRCA}}]\big].$$

We have

$$\mathbb{E}_{n,m}\big[\mathbb{V}_{N_{\tau_1},M_{\tau_1}}[T_{\mathrm{MRCA}}]\big] = \alpha_{n,m}v_{n-1,m} + \beta_{n,m}v_{n-1,m+1} + \gamma_{n,m}v_{n+1,m-1}$$

and

$$\mathbb{V}_{n,m}\big[\mathbb{E}_{N_{\tau_1},M_{\tau_1}}[T_{\mathrm{MRCA}}]\big] = \mathbb{E}_{n,m}\big[\mathbb{E}_{N_{\tau_1},M_{\tau_1}}[T_{\mathrm{MRCA}}]^2\big] - \mathbb{E}_{n,m}\big[\mathbb{E}_{N_{\tau_1},M_{\tau_1}}[T_{\mathrm{MRCA}}]\big]^2$$
$$= \alpha_{n,m}t_{n-1,m}^2 + \beta_{n,m}t_{n-1,m+1}^2 + \gamma_{n,m}t_{n+1,m-1}^2$$
$$- \big(\alpha_{n,m}t_{n-1,m} + \beta_{n,m}t_{n-1,m+1} + \gamma_{n,m}t_{n+1,m-1}\big)^2.$$

Combining the observations proves the result. □

Expectation and variance of the total tree length

Let $l_{n,m}^{(\mathrm{a})} := \mathbb{E}_{n,m}[L^{(\mathrm{a})}]$ and $l_{n,m}^{(\mathrm{d})} := \mathbb{E}_{n,m}[L^{(\mathrm{d})}]$ denote the expectations, and $w_{n,m}^{(\mathrm{a})} := \mathbb{V}_{n,m}[L^{(\mathrm{a})}]$ and $w_{n,m}^{(\mathrm{d})} := \mathbb{V}_{n,m}[L^{(\mathrm{d})}]$ the variances of the total tree lengths, and define the mixed second moment, $w_{n,m}^{(\mathrm{a},\mathrm{d})} := \mathbb{E}_{n,m}[L^{(\mathrm{a})}L^{(\mathrm{d})}]$.

Proposition S1.2 (Recursion: Total tree length). *For $n, m \in \mathbb{N}$ we have*

$$l_{n,m}^{(\mathrm{a})} = n\lambda_{n,m}^{-1} + \alpha_{n,m}l_{n-1,m}^{(\mathrm{a})} + \beta_{n,m}l_{n-1,m+1}^{(\mathrm{a})} + \gamma_{n,m}l_{n+1,m-1}^{(\mathrm{a})} \tag{S3}$$

$$l_{n,m}^{(\mathrm{d})} = m\lambda_{n,m}^{-1} + \alpha_{n,m}l_{n-1,m}^{(\mathrm{d})} + \beta_{n,m}l_{n-1,m+1}^{(\mathrm{d})} + \gamma_{n,m}l_{n+1,m-1}^{(\mathrm{d})}, \tag{S4}$$

and

$$w_{n,m}^{(\mathrm{a})} = n^2 \lambda_{n,m}^{-2} + \alpha_{n,m} w_{n-1,m}^{(\mathrm{a})} + \beta_{n,m} w_{n-1,m+1}^{(\mathrm{a})} + \gamma_{n,m} w_{n+1,m-1}^{(\mathrm{a})}$$
$$+ \alpha_{n,m} (l_{n-1,m}^{(\mathrm{a})})^2 + \beta_{n,m} (l_{n-1,m+1}^{(\mathrm{a})})^2 + \gamma_{n,m} (l_{n+1,m-1}^{(\mathrm{a})})^2$$
$$- \big(\alpha_{n,m} l_{n-1,m}^{(\mathrm{a})} + \beta_{n,m} l_{n-1,m+1}^{(\mathrm{a})} + \gamma_{n,m} l_{n+1,m-1}^{(\mathrm{a})} \big)^2, \qquad (\mathrm{S}5)$$

$$w_{n,m}^{(\mathrm{d})} = m^2 \lambda_{n,m}^{-2} + \alpha_{n,m} w_{n-1,m}^{(\mathrm{d})} + \beta_{n,m} w_{n-1,m+1}^{(\mathrm{d})} + \gamma_{n,m} w_{n+1,m-1}^{(\mathrm{d})}$$
$$+ \alpha_{n,m} (l_{n-1,m}^{(\mathrm{d})})^2 + \beta_{n,m} (l_{n-1,m+1}^{(\mathrm{d})})^2 + \gamma_{n,m} (l_{n+1,m-1}^{(\mathrm{d})})^2$$
$$- \big(\alpha_{n,m} l_{n-1,m}^{(\mathrm{d})} + \beta_{n,m} l_{n-1,m+1}^{(\mathrm{d})} + \gamma_{n,m} l_{n+1,m-1}^{(\mathrm{d})} \big)^2, \qquad (\mathrm{S}6)$$

$$w_{n,m}^{(\mathrm{a},\mathrm{d})} = 2nm \lambda_{n,m}^{-2} + \alpha_{n,m} w_{n-1,m}^{(\mathrm{a},\mathrm{d})} + \beta_{n,m} w_{n-1,m+1}^{(\mathrm{a},\mathrm{d})} + \gamma_{n,m} w_{n+1,m-1}^{(\mathrm{a},\mathrm{d})}. \qquad (\mathrm{S}7)$$

Proof of Proposition S1.2. The result can easily be obtained observing that each stretch of time of length τ in which we have a constant number of n active blocks and m dormant blocks contributes with $n\tau$ to the total active tree length, and with $m\tau$ to the total dormant tree length. Thus we have

$$l_{n,m}^{(\mathrm{a})} = n \mathbb{E}_{n,m}[\tau_1] + \mathbb{E}_{n,m} \big[\mathbb{E}_{N_{\tau_1}, M_{\tau_1}}[L^{(\mathrm{a})}] \big],$$

and we proceed as in the proof of Proposition S1.1. From these quantities we easily obtain the expected total tree length as $l_{n,m}^{(\mathrm{a})} + l_{n,m}^{(\mathrm{d})}$. Moreover,

$$\mathrm{Cov}_{n,m}(L^{(\mathrm{a})}, L^{(\mathrm{d})}) = w_{n,m}^{(\mathrm{a},\mathrm{d})} - w_{n,m}^{(\mathrm{a})} w_{n,m}^{(\mathrm{d})}.$$

$\square$

Expectation of total length of external branches

To derive recursions for the total length of external branches in either of the two states is a little more involved, since obviously a coalescence can happen between either two external active branches, two internal active branches, or an external and an internal active branch. We use indices (n, n', m, m') to denote the number of external active branches, internal active branches, external dormant branches, and internal dormant branches, respectively. Abbreviate

$$\alpha^{(1)}_{n,n',m,m'} := \frac{\binom{n}{2}}{\lambda_{n+n',m+m'}}, \quad \alpha^{(2)}_{n,n',m,m'} := \frac{\binom{n'}{2}}{\lambda_{n+n',m+m'}}, \quad \alpha^{(3)}_{n,n',m,m'} := \frac{nn'}{\lambda_{n+n',m+m'}},$$

$$\beta^{(1)}_{n,n'm,m'} := \frac{cn}{\lambda_{n+n',m+m'}}, \quad \beta^{(2)}_{n,n'm,m'} := \frac{cn'}{\lambda_{n+n',m+m'}},$$

$$\gamma^{(1)}_{n,n'm,m'} := \frac{cKm}{\lambda_{n+n',m+m'}}, \quad \gamma^{(2)}_{n,n',m,m'} := \frac{cKm'}{\lambda_{n+n',m+m'}}.$$

Let $E^{(\mathrm{a})}$ denote the total length of external branches in the plant state, and $E^{(\mathrm{d})}$ the total length of external branches in the seed state. Then we have

Proposition S1.3 (Recursion: Total length of external branches). *For $n, m \in \mathbb{N}$, we have the representation*

$$\mathbb{E}_{n,m}[E^{(\mathrm{a})}] = e^{(\mathrm{a})}_{n,0,m,0}, \quad \mathbb{E}_{n,m}[E^{(\mathrm{d})}] = e^{(\mathrm{d})}_{n,0,m,0},$$

where $e^{(\mathrm{a})}_{n,n',m,m'}$ and $e^{(\mathrm{d})}_{n,n',m,m'}$, $n, n', m, m' \in \mathbb{N}_0$ satisfy the recursions

$$\begin{aligned}
e^{(\mathrm{a})}_{n,n',m,m'} = {} & n\lambda^{-1}_{n+n',m+m'} \\
& + \alpha^{(1)}_{n,n',m,m'}e^{(\mathrm{a})}_{n-2,n'+1,m,m'} + \alpha^{(2)}_{n,n',m,m'}e^{(\mathrm{a})}_{n,n'-1,m,m'} + \alpha^{(3)}_{n,n',m,m'}e^{(\mathrm{a})}_{n-1,n',m,m'} \\
& + \beta^{(1)}_{n,n',m,m'}e^{(\mathrm{a})}_{n-1,n',m+1,m'} + \beta^{(2)}_{n,n',m,m'}e^{(\mathrm{a})}_{n,n'-1,m,m'+1} \\
& + \gamma^{(1)}_{n,n',m,m'}e^{(\mathrm{a})}_{n+1,n',m-1,m'} + \gamma^{(2)}_{n,n',m,m'}e^{(\mathrm{a})}_{n,n'+1,m,m'-1}
\end{aligned}$$

and

$$e_{n,n',m,m'}^{(d)} = m\lambda_{n+n',m+m'}^{-1}$$
$$+ \alpha_{n,n',m,m'}^{(1)} e_{n-2,n'+1,m,m'}^{(d)} + \alpha_{n,n',m,m'}^{(2)} e_{n,n'-1,m,m'}^{(d)} + \alpha_{n,n',m,m'}^{(3)} e_{n-1,n',m,m'}^{(d)}$$
$$+ \beta_{n,n',m,m'}^{(1)} e_{n-1,n',m+1,m'}^{(d)} + \beta_{n,n',m,m'}^{(2)} e_{n,n'-1,m,m'+1}^{(d)}$$
$$+ \gamma_{n,n',m,m'}^{(1)} e_{n+1,n',m-1,m'}^{(d)} + \gamma_{n,n',m,m'}^{(2)} e_{n,n'+1,m,m'-1}^{(d)}$$

Observing that $e_{0,n',0,m'}^{(a)} = e_{0,n',0,m'}^{(d)} = 0$ for all n', m', and $e_{1,0,0,0}^{(a)} = e_{1,0,0,0}^{(d)} = 0$, and that the total number $n+n'+m+m'$ is non-increasing, these recursions can be solved iteratively.

Proof of Proposition S1.3. This follows by a similar first-step analysis as in Proposition S1.2, taking into account the transitions for internal and external branches, and observing that at each coalescence event between two external branches, the number of external plant branches is reduced by two and the number of internal branches is increased by one, in a coalescence of an external and an internal branch, the number of external plant branches is reduced by one and the number of internal plant branches stays the same, and in a coalescence of two internal branches, their number is reduced by one. □

Obviously, the expected total length of external branches is then given by $e_{n,0,m,0}^{(a)} + e_{n,0,m,0}^{(d)}$. Note that proceeding as in Proposition S1.2, we could also give recursions for the variances of these quantities.

Expectation and variance of the number of segregating sites

Proposition S1.4. *For $n, m \in \mathbb{N}_0$ we have*

$$\mathbb{E}_{n,m}[S] = \frac{\theta_1}{2} l_{n,m}^{(a)} + \frac{\theta_2}{2} l_{n,m}^{(d)},$$

and

$$\mathbb{V}_{n,m}[S] = \frac{\theta_1}{2} l_{n,m}^{(a)} + \frac{\theta_2}{2} l_{n,m}^{(d)} + \frac{\theta_1^2}{4} w_{n,m}^{(a)} + \frac{\theta_2^2}{4} w_{n,m}^{(d)} + \frac{\theta_1 \theta_2}{2} (w_{n,m}^{a,d} - l_{n,m}^{(a)} l_{n,m}^{(d)}),$$

where $l_{n,m}^{(a)}, l_{n,m}^{(d)}, w_{n,m}^{(a)}, w_{n,m}^{(d)}$ and $w_{n,m}^{(a,d)}$ are given by Proposition S1.2.

Proof of Proposition S1.4. Observe that conditional on the total lengths $L^{(a)}, L^{(d)}$, the number of segregating sites is the sum of two independent Poisson random variables with parameters $\theta_1 L^{(a)}/2$ and $\theta_2 L^{(d)}/2$, respectively. Hence, if an ancestral line is in the plant state for a period of time of length $L > 0$, the expected number of mutations that occur in this period is $L\theta_1/2$. Similarly, in a period of length L when the ancestral line is a seed, the expected number of mutations is $L\theta_2/2$. Thus the first result follows directly from Proposition S1.2.

For the second result, we apply the law of total variance and obtain similarly that

$$
\begin{aligned}
\mathbb{V}_{n,m}(S) &= \mathbb{E}_{n,m}[\mathbb{V}(S \mid L^{(a)}, L^{(d)})] + \mathbb{V}_{n,m}(\mathbb{E}[S \mid L^{(a)}, L^{(d)}]) \\
&= \mathbb{E}_{n,m}\left[\frac{\theta_1}{2}L^{(a)} + \frac{\theta_2}{2}L^{(d)}\right] + \mathbb{V}_{n,m}\left(\frac{\theta_1}{2}L^{(a)} + \frac{\theta_2}{2}L^{(d)}\right) \\
&= \frac{\theta_1}{2}l_{n,m}^{(a)} + \frac{\theta_2}{2}l_{n,m}^{(d)} + \frac{\theta_1}{4}^2 w_{n,m}^{(a)} + \frac{\theta_2}{4}^2 w_{n,m}^{(d)} + 2\frac{\theta_1}{2}\frac{\theta_2}{2}\mathrm{Cov}_{n,m}(L^{(a)}, L^{(d)}).
\end{aligned}
$$

$\square$

It is possible to directly derive a recursion for the number of segregating sites without explicitly passing through calculating the tree lengths. Since it may be of use we state it here. Let

$$
s_{n,m} := \mathbb{E}_{n,m}[S], \quad \text{and} \quad z_{n,m} := \mathbb{V}_{n,m}(S).
$$

Proposition S1.5 (Alternative recursion). *Let* $n, m \in \mathbb{N}_0$. *Then*

$$
s_{n,m} = \left(\frac{\theta_1}{2}n + \frac{\theta_2}{2}m\right)\lambda_{n,m}^{-1} + \alpha_{n,m}s_{n-1,m} + \beta_{n,m}s_{n-1,m+1} + \gamma_{n,m}s_{n+1,m-1} \tag{S8}
$$

$$
\begin{aligned}
z_{n,m} = &\left(\frac{\theta_1}{2}n + \frac{\theta_2}{2}m\right)\lambda_{n,m}^{-1} + \left(\frac{\theta_1}{2}n + \frac{\theta_2}{2}m\right)^2\lambda_{n,m}^{-2} \\
&+ \alpha_{n,m}z_{n-1,m} + \beta_{n,m}z_{n-1,m+1} + \gamma_{n,m}z_{n+1,m-1} \\
&+ \alpha_{n,m}s_{n-1,m}^2 + \beta_{n,m}s_{n-1,m+1}^2 + \gamma_{n,m}s_{n+1,m-1}^2 \\
&- \left(\alpha_{n,m}s_{n-1,m} + \beta_{n,m}s_{n-1,m+1} + \gamma_{n,m}s_{n+1,m-1}\right)^2. \tag{S9}
\end{aligned}
$$

J. Blath *et al.*

Proof of Proposition S1.5. Let σ_1 denote the number of mutations that occur until time τ_1, which was defined in the proof of Proposition S1.1. Given $\tau_1 = t$, we know that σ_1 is the sum of two independent Poisson random variables with parameters $\theta_1 n t$ and $\theta_2 m t$, respectively. As in the previous proof we obtain

$$
\begin{aligned}
s_{n,m} &= \mathbb{E}_{n,m}[\sigma_1] + \mathbb{E}_{n,m}\big[\mathbb{E}_{N_{\tau_1},M_{\tau_1}}[S]\big] \\
&= \left(\frac{\theta_1}{2}n + \frac{\theta_2}{2}m\right)\mathbb{E}_{n,m}[\tau_1] + \alpha_{n,m}s_{n-1,m} + \beta_{n,m}s_{n-1,m+1} + \gamma_{n,m}s_{n+1,m-1}
\end{aligned}
$$

and

$$
z_{n,m} = \mathbb{V}_{n,m}(\sigma_1) + \mathbb{E}_{n,m}\big[\mathbb{V}_{N_{\tau_1},M_{\tau_1}}(S)\big] + \mathbb{V}_{n,m}\big(\mathbb{E}_{N_{\tau_1},M_{\tau_1}}[S]\big).
$$

Once more using the law of total variance we obtain

$$
\begin{aligned}
\mathbb{V}_{n,m}[\sigma_1] &= \mathbb{E}_{n,m}\big[\mathbb{V}_{n,m}(\sigma_1 \mid \tau_1)\big] + \mathbb{V}_{n,m}\big(\mathbb{E}_{n,m}[\sigma_1 \mid \tau_1]\big) \\
&= \left(\frac{\theta_1}{2}n + \frac{\theta_2}{2}m\right)\mathbb{E}_{n,m}[\tau_1] + \left(\frac{\theta_1}{2}n + \frac{\theta_2}{2}m\right)^2 \mathbb{V}_{n,m}[\tau_1] \\
&= \left(\frac{\theta_1}{2}n + \frac{\theta_2}{2}m\right)\lambda_{n,m}^{-1} + \left(\frac{\theta_1}{2}n + \frac{\theta_2}{2}m\right)^2 \lambda_{n,m}^{-2}. \tag{S10}
\end{aligned}
$$

The same calculations as in the proof of Proposition S1.1 lead to

$$
\mathbb{E}_{n,m}\big[\mathbb{V}_{N_{\tau_1},M_{\tau_1}}(S)\big] = \alpha_{n,m}z_{n-1,m} + \beta_{n,m}z_{n-1,m+1} + \gamma_{n,m}z_{n+1,m-1},
$$

and

$$
\begin{aligned}
\mathbb{V}_{n,m}\big(\mathbb{E}_{N_{\tau_1},M_{\tau_1}}[S]\big) &= \alpha_{n,m}s_{n-1,m}^2 + \beta_{n,m}s_{n-1,m+1}^2 + \gamma_{n,m}s_{n+1,m-1}^2 \\
&\quad - \big(\alpha_{n,m}s_{n-1,m} + \beta_{n,m}s_{n-1,m+1} + \gamma_{n,m}s_{n+1,m-1}\big)^2.
\end{aligned}
$$

$\square$

Expected value of average pairwise differences (π)

Recall the definition (19) of average pairwise difference $\pi = \binom{N_0+M_0}{2}^{-1} \sum_{(i,j): i<j} K_{i,j}$.

Proposition S1.6. *For $n, m \in \mathbb{N}_0$ we have*

$$\mathbb{E}_{n,m}\left[\pi\right] = \frac{1}{\binom{n+m}{2}}\left\{\binom{n}{2}\left(\frac{\theta_1}{2}l_{2,0}^{(a)} + \frac{\theta_2}{2}l_{2,0}^{(d)}\right) + nm\left(\frac{\theta_1}{2}l_{1,1}^{(a)} + \frac{\theta_2}{2}l_{1,1}^{(d)}\right) + \binom{m}{2}\left(\frac{\theta_1}{2}l_{0,1}^{(a)} + \frac{\theta_2}{2}l_{0,1}^{(d)}\right)\right\}$$

Proof of Proposition S1.6. By definition

$$\mathbb{E}_{n,m}\left[\pi\right] = \frac{1}{\binom{n+m}{2}}\sum_{1 \le i < j \le n+m}\mathbb{E}_{n,m}\left[K_{i,j}\right].$$

When compairing two individuals their *pairwise* differences in the infinite sites model coincide with the number of mutations that occured along the branches of their corresponding sub-tree and are thus given the product of the mutation rate and length of the branches. Therefore, $\mathbb{E}_{n,m}[K_{i,j}]$ actually only depends on whether i, j are dormant or active individuals. We obtain

$$\mathbb{E}_{n,m}\left[K_{i,j}\right] = \begin{cases} \frac{\theta_1}{2}l_{2,0}^{(a)} + \frac{\theta_2}{2}l_{2,0}^{(d)}, & \text{if } i, j \text{ are active} \\ \frac{\theta_1}{2}l_{1,1}^{(a)} + \frac{\theta_2}{2}l_{1,1}^{(d)}, & \text{if } i \text{ is active and } j \text{ dormant} \\ \frac{\theta_1}{2}l_{0,2}^{(a)} + \frac{\theta_2}{2}l_{0,2}^{(d)}, & \text{if } i, j \text{ are dormant.} \end{cases}$$

Substituting this into the above equation, the result follows. $\qquad\square$

File S2   Solving the recursions numerically

Since all the recursions have the same general form, a generic method for solving them numerically will now be given. The idea is to use standard linear algebra methods to solve the standard linear system $A\underline{t} = \underline{b}$. Let $\underline{t} = (t_0, t_1, \ldots, t_{n+m})$ denote the vector of quantities we are solving for, where we order them according to number of active lines. For any given number n of active blocks and m of inactive blocks, so the current total number of blocks is $n + m$, write $\underline{t} = (t_0, t_1, \ldots, t_{n+m})$ where $t_i \equiv t_{i,n+m-i}$, and write $\ell := n + m$. Let $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ denote square $(\ell+1) \times (\ell+1)$ matrices whose rows and columns are enumerated from 0; with non-zero terms $a_{i-1,i} = \alpha_{i,\ell-i}$, $b_{i,i-1} = \beta_{i,\ell-i}$, $c_{i,i+1} = \gamma_{i,\ell-i}$, and let $\boldsymbol{I}$ denote a $(\ell+1) \times (\ell+1)$ identity matrix. Assume, by way of example, we are solving the recursion (10) for expected time to most recent common ancestor. Define the vector $\underline{k}$ with elements $k_i = 1/\lambda_{i,\ell-i}$, and $\underline{r} = (0, r_0, r_1, \ldots, r_{n+m-1})$ where $r_j = t_{j,n+m-j-1}$.

The recursion in Proposition (S1.1) can now be written

$$\underline{t} = \underline{k} + \boldsymbol{A}\underline{r} + (\boldsymbol{B} + \boldsymbol{C})\underline{t}$$

Assuming we solve for $\underline{t}$ iteratively, starting from $n+m = 2$, $\underline{r}$ is a vector of known constants; hence

$$\underline{s} = (\boldsymbol{I} - \boldsymbol{B} - \boldsymbol{C})^{-1}(\underline{k} + \boldsymbol{A}\underline{r})$$

where $\boldsymbol{I} - \boldsymbol{B} - \boldsymbol{C}$ should be non-singular and $(\boldsymbol{I} - \boldsymbol{B} - \boldsymbol{C})^{-1}$ easily computable.

Similar methods may be applied to the other recursions.

Table S1: Relative expected lengths of external branches. The relative expected lengths of external branches $e_{(\underline{n})}^{(a)} / \left(e_{(\underline{n})}^{(a)} + e_{(\underline{n})}^{(d)} \right)$ from Prop. S1.3 with sample configuration $\underline{n} = (10, 0, 10, 0)$.

	$K = 0.01$, values of d					
c	0.001	0.01	0.1	1	10	100
0.001	0.0161	0.0449	0.0426	0.0195	0.0869	0.338
0.01	0.00973	0.0161	0.0457	0.0521	0.0928	0.338
0.1	0.00914	0.00987	0.0166	0.0519	0.116	0.335
1	0.00955	0.00963	0.0104	0.0183	0.0756	0.289
10	0.00985	0.00986	0.00995	0.0108	0.0194	0.095
100	0.0099	0.0099	0.00991	0.00999	0.0109	0.0196
	$K = 1$, values of d					
c	0.001	0.01	0.1	1	10	100
0.001	0.0313	0.02	0.0876	0.338	0.725	0.954
0.01	0.0516	0.0499	0.0984	0.339	0.725	0.954
0.1	0.116	0.12	0.155	0.349	0.723	0.953
1	0.287	0.289	0.301	0.396	0.704	0.946
10	0.449	0.45	0.452	0.471	0.605	0.884
100	0.494	0.494	0.494	0.496	0.517	0.659
	$c = 1$, values of d					
K	0.001	0.01	0.1	1	10	100
0.001	0.000996	0.001	0.00109	0.00198	0.0104	0.071
0.01	0.00955	0.00963	0.0104	0.0183	0.0756	0.289
0.1	0.0705	0.071	0.0756	0.115	0.301	0.689
1	0.287	0.289	0.301	0.396	0.704	0.946
10	0.687	0.689	0.704	0.797	0.95	0.994
100	0.945	0.946	0.95	0.971	0.995	0.999
	$c = 0.01$, values of d					
K	0.001	0.01	0.1	1	10	100
0.001	0.00109	0.00195	0.00973	0.0442	0.0522	0.0928
0.01	0.00973	0.0161	0.0457	0.0521	0.0928	0.338
0.1	0.0457	0.0539	0.0516	0.0934	0.338	0.725
1	0.0516	0.0499	0.0984	0.339	0.725	0.954
10	0.0984	0.143	0.352	0.726	0.954	0.995
100	0.352	0.449	0.74	0.954	0.995	0.999

Table S2: Expected length of external branches. The expected total lengths of external branches $\left(e_{(\underline{n})}^{(\mathrm{a})}, e_{(\underline{n})}^{(\mathrm{d})}\right)$ from Prop. S1.3 with sample configuration $\underline{n} = (10, 0, 0, 0)$, as a function of c and K. The expected length $e_{(n)} = 2$ when associated with the Kingman coalescent (FU, 1995).

|	$c = 1$, values of d		
K	0.001	1	100
0.001	(1.22e+03, 1.22e+06)	(610, 3.05e+05)	(14.2, 141)
0.01	(124, 1.24e+04)	(63, 3.15e+03)	(3.4, 3.36)
0.1	(14.3, 143)	(8.28, 41.4)	(2.18, 0.216)
1	(3.41, 3.4)	(2.77, 1.39)	(2.02, 0.02)
10	(2.18, 0.218)	(2.1, 0.105)	(2, 0.00198)
100	(2.02, 0.0202)	(2.01, 0.01)	(2, 0.000198)
	$K = 0.01$, values of d		
c	0.001	1	100
0.001	(56.7, 2.83e+03)	(2.03, 0.203)	(2, 0.002)
0.01	(102, 9.28e+03)	(2.68, 2.65)	(2.01, 0.0201)
0.1	(111, 1.1e+04)	(11.5, 104)	(2.12, 0.211)
1	(124, 1.24e+04)	(63, 3.15e+03)	(3.4, 3.36)
10	(174, 1.74e+04)	(158, 1.44e+04)	(17.9, 163)
100	(198, 1.98e+04)	(196, 1.94e+04)	(100, 5.01e+03)

J. Blath *et al.*
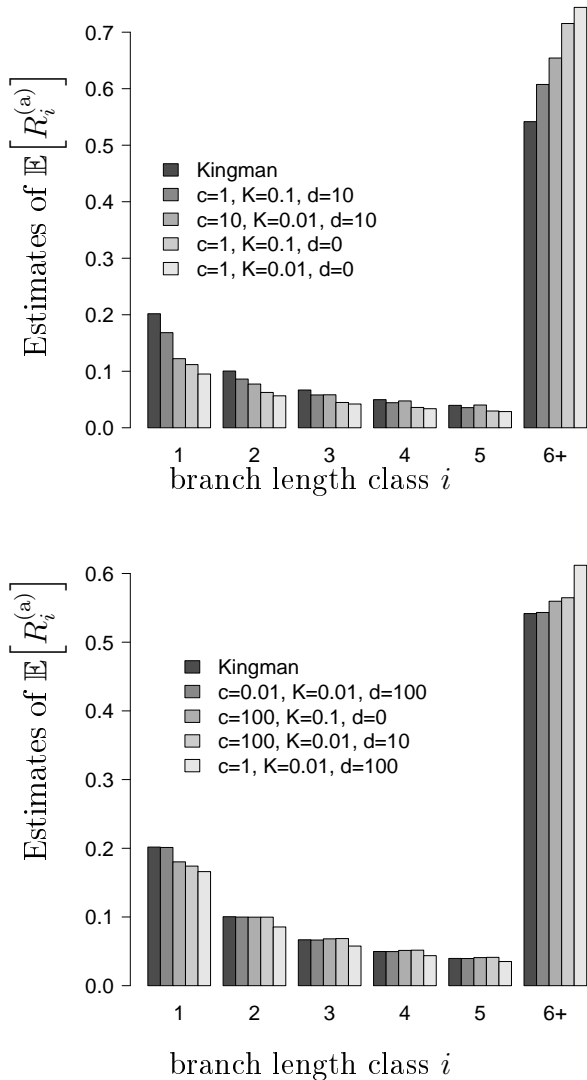
Figure S1: Expected normalised branch lengths. Estimates of the expected normalized branch lengths $\mathbb{E}\left[R_i^{(a)}\right]$, with $R_i^{(a)} := \frac{B_i^{(a)}}{B^{(a)}}$ with $B_i^{(a)}$ denoting the random total length of *active* branches subtending i leaves, and $B^{(a)}$ the sum of $B_i^{(a)}$; with all $n = 100$ sampled lines assumed active, and values of c, K, d as shown. The values labelled **6+** denote the collected tail $\overline{R}_6^{(a)} + \cdots + \overline{R}_{99}^{(a)}$. All estimates based on 10^5 replicates.
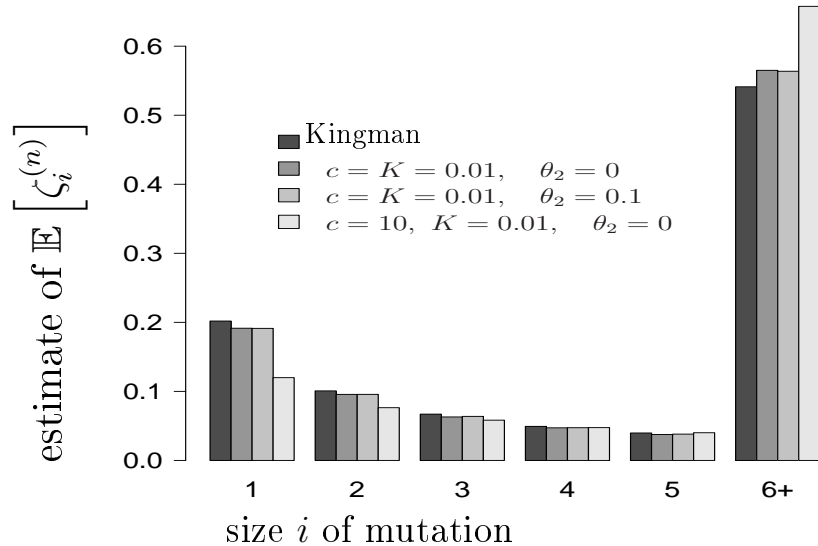
Figure S2: Expected normalised site-frequency spectrum. Estimates $\overline{\zeta}_i^{(n)}$ $\left(\zeta_i^{(n)} = \frac{\xi_i^{(n)}}{|\xi^{(n)}|}\right)$ where $|\xi^{(n)}| = \xi_1^{(n)} + \cdots + \xi_{n-1}^{(n)}$ denotes the total number of segregating sites, of expected normalized spectra $\mathbb{E}\left[\zeta_i^{(n)}\right]$ with all $n = 100$ sampled lines assumed active, active mutation rate $\theta_1 = 2$, and with c, K, and inactive mutation rate θ_2 as shown. The entries labelled '6+' represent the collected tail $\overline{\zeta}_{6+}^{(n)} = \sum_{i \geq 6} \overline{\zeta}_i^{(n)}$. Estimates are based on 10^5 replicates.
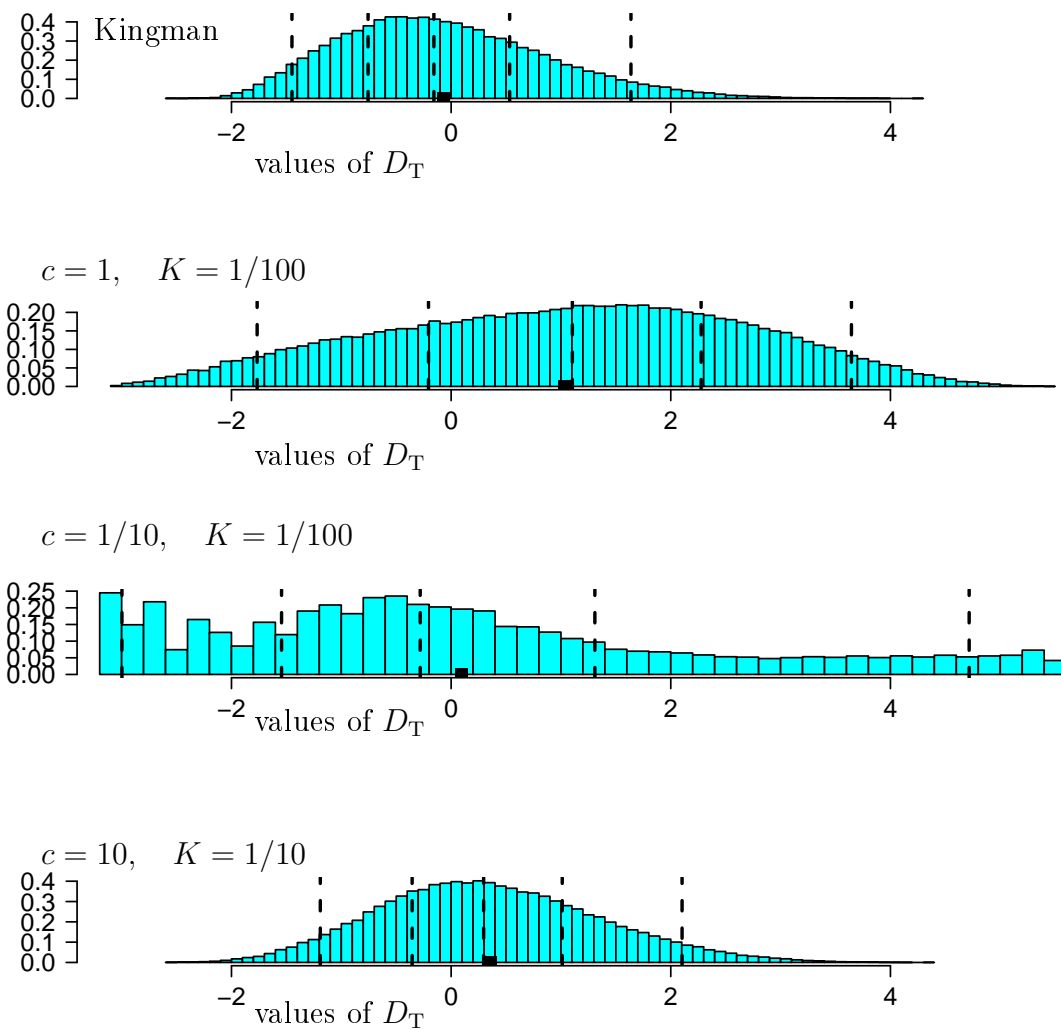
Figure S3: Tajima's statistic D_T (24). Estimates of the distribution of Tajima's statistic D_T (24) with all $n = 100$ sampled lines assumed active, $\theta_1 = 2$, $\theta_2 = 0$. The vertical broken lines are the 5%, 25%, 50%, 75%, 95% quantiles and the black square (■) denotes the mean. The entries are normalised to have unit mass 1. The histograms are drawn on the same horizontal scale. Based on 10^5 replicates.