CrossMark

J. Appl. Probab. **59,** 631–651 (2022)
*doi:*10.1017/jpr.2021.79

THE SHAPE OF A SEED BANK TREE

ADRIÁN GONZÁLEZ CASANOVA,* ** *Universidad Nacional Autónoma de México*
LIZBETH PEÑALOZA,*** **** *Universidad del Mar*
ARNO SIRI-JÉGOUSSE,* ***** *Universidad Nacional Autónoma de México*

Abstract

We derive the asymptotic behavior of the total, active, and inactive branch lengths of the seed bank coalescent when the initial sample size grows to infinity. These random variables have important applications for populations evolving under some seed bank effects, such as plants and bacteria, and for some cases of structured populations like metapopulations. The proof relies on the analysis of the tree at a stopping time corresponding to the first time a deactivated lineage is reactivated. We also give conditional sampling formulas for the random partition, and we study the system at the time of the first reactivation of a lineage. All these results provide a good picture of the different regimes and behaviors of the block-counting process of the seed bank coalescent.

Keywords: Seed bank; structured coalescent; branch lengths; sampling formula

2020 Mathematics Subject Classification: Primary 60J95
Secondary 60C05; 60J28; 92D25

1. Introduction

Seeds, cysts, and other forms of dormancy generate seed banks, which store genetic information that can be temporarily lost from a population at a certain time and *resuscitate* later. Having a seed bank is a prevalent evolutionary strategy which has important consequences. For example, in the case of bacteria, it buffers against the selective pressure caused by environmental variability, and at the same time increases genetic variation [10, 15, 17].

A first attempt to construct a probabilistic model to study this phenomenon was [12], which considered a modified Wright–Fisher model in which each individual *chooses* its parent from the individuals at several generations in the past, and not only from the previous one. This construction has an important technical complication arising from the loss of the Markov property. A new model was defined and studied in [1] to avoid this issue. It consists of a two-level discrete Markov chain, which again generalizes the Wright–Fisher model.

Consider a haploid population of fixed size N which supports a seed bank of constant size M. The N *active* individuals are called *plants*, and the M *dormant* individuals are called *seeds*. Let $0 \leq \varepsilon \leq 1$ be such that $\lfloor \varepsilon N \rfloor \leq M$. The N plants from generation 0 produce N individuals in generation 1 by multinomial sampling (as in the Wright–Fisher model). However, $N - \lfloor \varepsilon N \rfloor$

Received 24 September 2020; revision received 30 August 2021.
* Postal address: Av. Universidad 3000, Col. UNAM Ciudad Universitaria, Delegación Coyoacán C.P. 04510 Ciudad de México.
** Email address: adrian.gonzalez@im.unam.mx
*** Postal address: Ciudad Universitaria, Santa María Huatulco, Oaxaca, México C.P. 70989.
**** Email address: lizbethpv@aulavirtual.umar.mx
***** Email address: arno@sigma.iimas.unam.mx

© The Author(s), 2022. Published by Cambridge University Press on behalf of Applied Probability Trust.

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

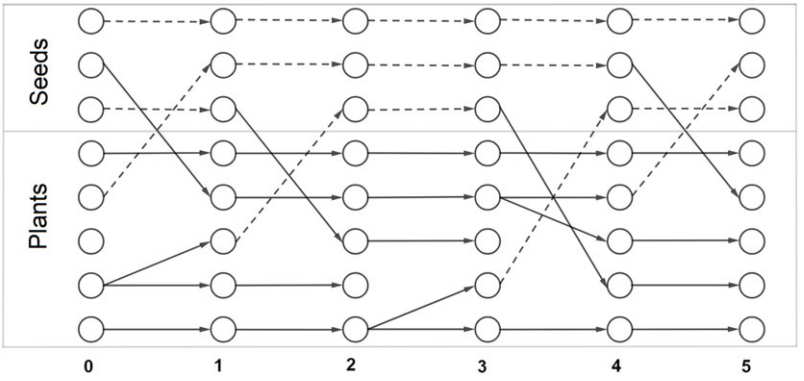


FIGURE 1. The discrete seed bank model. In this picture $N = 5$, $M = 3$, and $\lfloor \varepsilon N \rfloor = 1$, i.e. in each generation four plants are produced by active individuals, one seed germinates, and one plant creates one inactive individual.

randomly chosen of these individuals are plants and $\lfloor \varepsilon N \rfloor$ are seeds. Then, $\lfloor \varepsilon N \rfloor$ uniformly (without replacement) sampled seeds from the seed bank in generation 0 become plants in generation 1. The $\lfloor \varepsilon N \rfloor$ seeds produced by the plants in generation 1 take the place of the seeds that germinate. Thus, we again have N plants and M seeds in generation 1 (see Figure 1). This random mechanism is repeated independently to produce the next generations. Observe that this model, unlike [12], has non-overlapping reproduction events.

If we let N (and M) go to infinity and rescale the time, the stochastic process that describes the limiting gene genealogy of a sample taken from the seed bank model is called the *seed bank coalescent* [1]. Apart from populations of plants or bacteria, it is remarkable that the seed bank coalescent is a convenient genealogical model for some metapopulations [14]. In fact, it was independently introduced in that context and named the *peripatric coalescent*. It corresponds to a special modification of structured coalescence in which small colonies can emerge from a main population and merge again with it. The seed bank coalescent is a structured coalescent with an active part, having the dynamics of a *Kingman coalescent*, and a dormant part where the lineages are frozen. Lineages can activate or deactivate at certain rates; see Figure 2 for an illustration.

In this paper we study the asymptotic behavior of some functionals of the seed bank tree that shed light on connections between theoretical and applied population genetics. As an illustration, there is a close relation between the shape of the genealogical tree of a sample of size n and the number of mutations observed in it. More precisely, suppose that mutations appear in the genealogy by simply superimposing a Poisson process on the ancestral lineages (as in the infinite sites model [7, Section 1.4]). Then, the shape of the tree determines the distribution of the data obtained by DNA sequencing, and it can therefore be inferred from it. For example, conditionally on the total length of the coalescent, denoted by L_n, the number of mutations observed in the sample has a Poisson distribution with parameter μL_n, where μ is the mutation rate. Thus, if we know the asymptotic behavior of the total length of the tree, we can deduce the asymptotic behavior of the number of mutations. This is the key tool for obtaining a Watterson-type estimator for the mutation rate, see [7]. Not surprisingly, the asymptotics of the total length of many classical coalescents have been studied, e.g. in [5, 2, 13, 4].

It was established in [1] that the time to the most recent common ancestor of a sample of size n in the seed bank coalescent is of order $\log \log n$. This is an important difference from

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

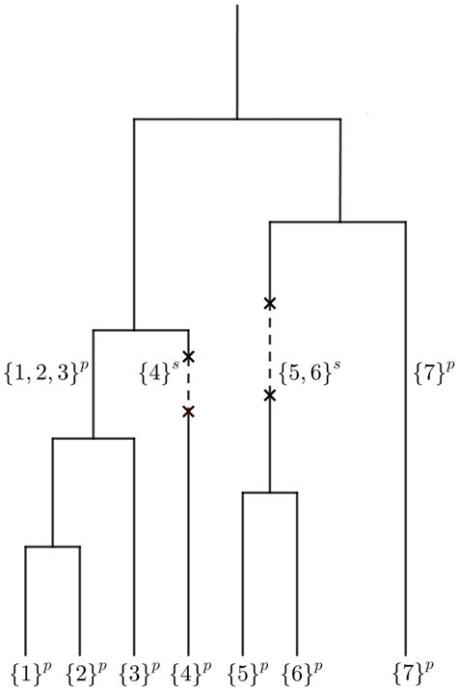


FIGURE 2. A possible realization of the seed bank 7-coalescent. Dotted lines indicate inactive individuals and the crosses mean that there is a deactivation or a reactivation.

the classical Kingman coalescent, whose height is finite. Our study establishes that the total length of the tree built from a sample of n plants and zero seeds is of the same order as that of the Kingman coalescent, behaving like $\log n$, but with a different multiplicative constant depending on the activation and deactivation parameters of the model. Moreover, we show that the total active length behaves like the total length of the Kingman coalescent. This means that it is not possible to distinguish between the null Kingman model and the alternative seed bank model using only the tree length (which can be inferred from the number of mutations) unless the dormant individuals have the possibility to mutate while being in the seed bank; that is actually the case in the metapopulation framework described in [14]. This conclusion agrees with the main result in [16], where Maughan observed experimentally that a population of bacteria undergoing dormancy typically does not have a significantly different number of mutations than a population of bacteria not undergoing dormancy. Furthermore, our results offer new insights on the reason for this: most of the mutations occur in the *Kingman phase* (shortly before the leaves), and in this part of the ancestral tree, dormancy is irrelevant. On the other hand, populations suffering a significant number of mutations while being in the dormant state would be expected to have a higher evolutionary rate. This remark together with [16] suggests that the mutations that occur to individuals in the latent state are atypical. This is opposed to previous works suggesting that the normal rate of molecular evolution of bacteria with a seed bank is evidence that mutations affecting dormant individuals are frequent [16].

More theoretical and experimental work is needed to clarify the role of dormancy in the flow of evolution. Finer results, such as sampling formulas, can be derived to discriminate between both null and seed bank models. For the time being, we are able to describe the seed bank tree

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

TABLE 1. Summary of the asymptotic behavior of the functionals of the seed bank coalescent studied in this work. Here, Y is a Beta$(2c_1, 1)$ distributed random variable, T is an exponential random variable with parameter $2c_1c_2$, and Z is a Fréchet random variable with shape parameter 1 and scale parameter $4c_1c_2$.

Stopping time (τ)	Asymptotics of τ	Asymptotics of $N_n(\tau)$	Asymptotics of $M_n(\tau)$
γ_n	$2(1-Y)/Yn$	Yn	1
θ_n	$T/\log n$	$Z \log n$	$2c_1 \log n$
σ_n	$\log\log n$	1	0

in detail as it undergoes different phases. It can be said that we describe the shape of the seed bank tree.

1.1. Main results

We study some relevant stopping times of the seed bank coalescent, leading to a complete description of the shape of the tree and explaining how long the genealogies spend in successive dynamical phases, as detailed precisely in Table 1 and Figure 3.

Let us now define the seed bank coalescent properly. Fix $n \in \mathbb{N}$ and let $\mathcal{P}_n$ be the set of partitions of $[n] := \{1, 2, \ldots, n\}$. Then the set of marked partitions $\mathcal{P}_n^{\{\mathrm{p},\mathrm{s}\}}$ is constructed from $\mathcal{P}_n$ by adding a flag (either p for a plant or s for a seed) to each block of the partition. For example, for $n = 7$, $\pi = \{\{1, 2, 3\}^{\mathrm{p}}, \{4\}^{\mathrm{s}}, \{5, 6\}^{\mathrm{s}}, \{7\}^{\mathrm{p}}\}$ is an element of $\mathcal{P}_7^{\{\mathrm{p},\mathrm{s}\}}$. The seed bank n-coalescent $(\Pi_n(t))_{t\geq 0}$, with deactivation intensity $c_1 > 0$ and activation intensity $c_2 > 0$, is the continuous-time Markov chain with values in $\mathcal{P}_n^{\{\mathrm{p},\mathrm{s}\}}$ having the following dynamics. As for the Kingman coalescent, each pair of plant blocks merges at rate 1, independent of each other. Moreover, any block can change its flag from p to s at rate c_1, and vice versa at rate c_2; see Figure 2 for an illustration.

The block-counting process of the seed bank n-coalescent is the two-dimensional Markov chain $\big(N_n(t), M_n(t)\big)_{t\geq 0}$ with values in $([n] \cup \{0\}) \times ([n] \cup \{0\})$ and the following transition rates, for $t \geq 0$:

$$\big(N_n(t), M_n(t)\big) \text{ jumps from } (i,j) \text{ to } \begin{cases} (i-1, j) & \text{at rate } \binom{i}{2} \quad \text{(coalescence)}, \\ (i-1, j+1) & \text{at rate } c_1 i \quad \text{(deactivation)}, \\ (i+1, j-1) & \text{at rate } c_2 j \quad \text{(activation)}. \end{cases}$$

Note that $N_n(t)$ can have either an upward jump if a seed becomes a plant, or a downward jump if there is a coalescence event or a plant becomes a seed. Also observe that each jump has size one. In the following, we suppose that $N_n(0) = n$ and $M_n(0) = 0$.

For $i \in [n]$, we denote by τ_n^i the hitting time of level i by the process N_n, i.e. $\tau_n^n = 0$ and

$$\tau_n^i = \inf\{t \geq 0 : N_n(t) = i\}. \tag{1.1}$$

Furthermore, let γ_n and θ_n be, respectively, the first time that some plant becomes a seed and the first time that some seed becomes a plant, i.e.

$$\gamma_n = \inf\{t > 0 : M_n(t-) < M_n(t)\} = \inf\{t > 0 : M_n(t) = 1\}, \tag{1.2}$$

$$\theta_n = \inf\{t > 0 : M_n(t-) > M_n(t)\}. \tag{1.3}$$

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

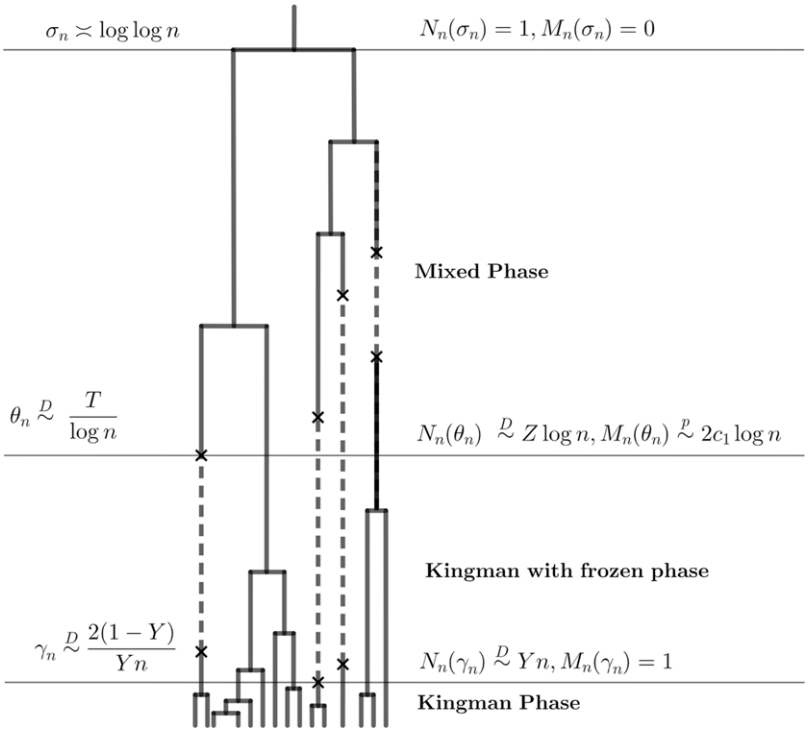


FIGURE 3. Summary of the asymptotic behavior of the functionals of the seed bank coalescent studied in this work. Here, Y is a Beta$(2c_1, 1)$ distributed random variable, T is an exponential random variable with parameter $2c_1c_2$, and Z is a Fréchet random variable with shape parameter 1 and scale parameter $4c_1c_2$. The symbol $A_n \overset{\mathrm{p}}{\sim} B_n$ means that $\frac{A_n}{B_n} \to 1$ in probability. The symbol $A_n \overset{\mathcal{D}}{\sim} XB_n$ means that $\frac{A_n}{B_n} \to X$ in distribution. The symbol $A_n \asymp B_n$ means that $C_1B_n \le \mathbb{E}[A_n] \le C_2B_n$ for some constants C_1, C_2.

Finally, denote by σ_n the time to the most recent common ancestor, already studied in [1], $\sigma_n = \inf\{t > 0 : N_n(t) + M_n(t) = 1\} = \inf\{t > 0 : N_n(t) = 1, M_n(t) = 0\}$.

We first obtain asymptotic results on the random variables γ_n and θ_n and the size of the system at those times. The results obtained in Sections 2 and 3 are summarized in Table 1 and Figure 3.

Observe that the rate of coalescence is quadratic with respect to the number of plants, while the rate of deactivation (resp. the rate of activation) is linear with respect to the number of plants (resp. the number of seeds). From [1], we inferred that the number of seeds accumulated until time θ_n, $M_n(\theta_n)$, is of order $\log n$. Lemma 3.3 suggests that, until time $\tau_n^{\lfloor(\log n)^a\rfloor}$, for $a > 1/2$, the block-counting process $(N_n(t))_{t\ge0}$ behaves similarly to that of the Kingman coalescent. However, at time $\tau_n^{\lfloor\sqrt{\log n}\rfloor}$, the system reaches a level of $\sqrt{\log n}$ plants and the times of decay are no longer close to those of the Kingman coalescent. Indeed, at this time, we argue that the number of seeds is still of order $\log n$ and the coalescence events no longer dominate the dynamics. The seed bank coalescent then enters a mixed regime with coalescence and activation occurring at the same velocity.

In Section 4 we analyze the total length

$$L_n = A_n + I_n, \tag{1.4}$$

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

where the *active* and *inactive* lengths are defined by

$$A_n = \int_0^{\sigma_n} N_n(t)\,\mathrm{d}t, \qquad I_n = \int_0^{\sigma_n} M_n(t)\,\mathrm{d}t. \tag{1.5}$$

Our main result is stated as follows.

Theorem 1.1. *Consider the seed bank coalescent starting with n plants and no seeds. Then,*

$$\lim_{n\to\infty} \frac{L_n}{\log n} = 2\left(1 + \frac{c_1}{c_2}\right) \quad \textit{in probability.}$$

Interestingly, the numerical techniques of [11] *used to study the total length for fixed n show that the balance between active and inactive lengths is equally conserved for their expectations for any* $n \geq 2$, $c_1\mathbb{E}[A_n] = c_2\mathbb{E}[I_n]$. *The behavior of both* A_n *and* I_n *is obtained by considering those variables before and after the time of the first activation* θ_n. *Hence, the results of Section 3 are key tools for the forthcoming proofs. Theorem 1.1 also gives an immediate corollary on the number of active and inactive mutations on the seed bank tree.*

Corollary 1.1. *Consider the seed bank coalescent starting with n plants and no seeds. Let* S_n *be the number of mutations in the seed bank tree. Let* μ *be the mutation rate for the active individuals and let* κ *be the mutation rate for the inactive individuals. Then*

$$\lim_{n\to\infty} \frac{S_n}{\log n} = 2\left(\mu + \kappa\frac{c_1}{c_2}\right) \quad \textit{in probability.}$$

Finally, in Section 5 we establish a sampling formula inspired by Watterson's ideas in [18], which helps us to understand the fine configuration of the blocks of a seed bank coalescent at given times.

2. The time of the first deactivation

We start with the study of γ_n, the time of the first deactivation defined in (1.2), and the size of the system at this time. Observe that if $N_n(0) = n$ and $M_n(0) = 0$, there are $n - N_n(\gamma_n) - 1$ coalescence events until time γ_n and we can write $\gamma_n = \sum_{i=N_n(\gamma_n)+1}^{n} V_i$, where the V_i are independent exponential random variables with respective parameters $\binom{i}{2} + c_1 i$.

We start with an easy limit result on the variable $N_n(\gamma_n)$. Note that, in a classical Kingman coalescent with mutations appearing at rate c_1, the quantity $n - N_n(\gamma_n) - 1$ can also be interpreted as the number of coalescence events before the most recent mutation in the genealogy. Recent studies on the shape of coalescent trees at the time of the first mutation in a branch can be found in [8], with some direct applications to coalescent model selection [9].

Proposition 2.1. *Consider a seed bank coalescent starting with n plants and no seeds. Then* $\lim_{n\to\infty}\{N_n(\gamma_n)/n\} = Y$ *in distribution, where* $Y \sim \mathrm{Beta}(2c_1, 1)$.

Proof. Let $z \in (0, 1)$. We have

$$\mathbb{P}(N_n(\gamma_n) \leq zn) = \prod_{i=\lfloor zn\rfloor+1}^{n} \frac{\binom{i}{2}}{\binom{i}{2} + c_1 i} = \prod_{i=\lfloor zn\rfloor}^{n-1} \frac{i}{i + 2c_1} = \exp\left\{-\sum_{i=\lfloor zn\rfloor}^{n-1} \log\left(1 + \frac{2c_1}{i}\right)\right\}.$$

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

Using that $\log(1+x) \sim x$ near 0, we obtain

$$\mathbb{P}(N_n(\gamma_n) \le zn) \sim \exp\left\{ -\sum_{i=\lfloor zn \rfloor}^{n-1} \frac{2c_1}{i} \right\} \sim \exp\left\{ -2c_1 \log\left(\frac{1}{z}\right) \right\} = z^{2c_1},$$

which is the distribution function of a Beta$(2c_1, 1)$ random variable. □

We now establish the asymptotic behavior of the time of the first deactivation, γ_n.

Proposition 2.2. *Consider a seed bank coalescent starting with n plants and no seeds. Then,*

$$\lim_{n\to\infty} n\gamma_n = \Gamma := \frac{2(1-Y)}{Y} \tag{2.1}$$

in distribution, where Y is Beta$(2c_1, 1)$ *distributed. The density function of* Γ *is*

$$f_\Gamma(x) = c_1 \left(\frac{2}{2+x}\right)^{2c_1+1}$$

for $x \ge 0$. *In particular, if* $c_1 > 1/2$, *then the expectation of* Γ *is finite,*

$$\mathbb{E}[\Gamma] = \frac{2}{2c_1 - 1},$$

and if $c_1 > 1$, *the variance of* Γ *is finite,*

$$\mathrm{Var}(\Gamma) = \frac{4c_1}{(c_1-1)(2c_1-1)^2}.$$

Proof. Let $G_n(0) = 0$ and, for $t \in (0, 1)$, define

$$G_n(t) = \sum_{i=\lfloor (1-t)n \rfloor + 1}^{n} V_i = \sum_{i=\lfloor (1-t)n \rfloor + 1}^{n} \frac{2e_i}{i(i-1+2c_1)},$$

where the e_i are independent and identically distributed (i.i.d.) standard exponential random variables. With this notation, we obtain $\gamma_n = G_n(1 - N_n(\gamma_n)/n)$.

We first show that, for any $t \in (0, 1)$, we have

$$\lim_{n\to\infty} (nG_n(s))_{s\le t} = \left(\frac{2s}{1-s}\right)_{s\le t} \tag{2.2}$$

in distribution, in the sense of weak convergence in the path space D[0, t]. To this end, we first establish that, for a fixed $t \in (0, 1)$,

$$\lim_{n\to\infty} nG_n(t) = \frac{2t}{1-t} \tag{2.3}$$

in L^2. By definition, we have that

$$\mathbb{E}[nG_n(t)] = \sum_{i=\lfloor (1-t)n \rfloor + 1}^{n} \frac{2n}{i(i-1+2c_1)} \sim \frac{1}{n} \sum_{i=\lfloor (1-t)n \rfloor + 1}^{n} \frac{2}{(i/n)^2}.$$

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

By a Riemann sum argument, we obtain

$$\mathbb{E}[nG_n(t)] \sim \int_{1-t}^{1} \frac{2}{x^2}\,\mathrm{d}x = \frac{2t}{1-t}.$$

Now, by the independence of the random variables e_i,

$$\mathrm{Var}(nG_n(t)) = \sum_{i=\lfloor(1-t)n\rfloor+1}^{n} \frac{4n^2}{i^2(i-1+2c_1)^2} \sim \sum_{i=\lfloor(1-t)n\rfloor+1}^{n} \frac{4n^2}{i^4}.$$

Again, by a Riemann sum argument we obtain that $\mathrm{Var}(nG_n(t))$ converges to 0 as $n \to \infty$. This gives (2.3).

To obtain (2.2) we follow the same steps as [3, Proposition 6.1], with $\alpha = 2$. Then, the proof of (2.1) follows by adapting the alternative proof of [3, Theorem 5.2, p. 1713], taking $\alpha = 2$ and the limit variable σ being $1 - Y$ and Beta$(1, 2c_1)$ distributed.

The distribution function of Γ is given by

$$\mathbb{P}(\Gamma \le x) = \mathbb{P}\left(Y \ge \frac{2}{2+x}\right) = 1 - \left(\frac{2}{2+x}\right)^{2c_1}, \quad x \ge 0.$$

We get the density by differentiating. The moments of Γ are obtained by computing

$$\mathbb{E}\left[\Gamma^k\right] = \int_0^\infty kx^{k-1}\mathbb{P}(\Gamma > x)\,\mathrm{d}x = \int_0^\infty kx^{k-1}\left(\frac{2}{2+x}\right)^{2c_1}\mathrm{d}x.$$

In particular, the kth moment is finite for $c_1 > k/2$. □

3. The time of the first activation

In this section we study θ_n, the first time that a seed becomes a plant, which we introduced in (1.3). We also provide some limit laws for $N_n(\theta_n)$ and $M_n(\theta_n)$. Observe that from time zero up to time θ_n only two types of event occur, either coalescence or deactivation. Recall the successive hitting times of the chain N_n, denoted by $\left(\tau_n^i\right)_{i=1}^n$ and defined in (1.1).

Proposition 3.1. *Consider a seed bank coalescent starting with n plants and no seeds. Then the following asymptotics hold:*

$$\lim_{n\to\infty} \frac{N_n(\theta_n)}{\log n} = Z \quad \textit{in distribution}, \tag{3.1}$$

where Z is a Fréchet random variable with shape parameter 1 and scale parameter $4c_1c_2$*, with distribution function* $\mathbb{P}(Z \le z) = \exp\{-4c_1c_2/z\}$*. Also,*

$$\lim_{n\to\infty} \frac{M_n(\theta_n)}{\log n} = 2c_1 \quad \textit{in probability}. \tag{3.2}$$

Finally,

$$\lim_{n\to\infty} \log n\theta_n = T \quad \textit{in distribution}, \tag{3.3}$$

where T is an exponential random variable with parameter $2c_1c_2$*.*

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

The proof of (3.2) is obtained by combining Lemmas 3.1 and 3.4 The proof of (3.1) and (3.3) is obtained by combining Lemmas 3.2 and 3.5, which appear in the following. We get these results by coupling the seed bank coalescent with two simpler models.

The *colored* seed bank coalescent (see [1, Definition 4.2]) is a marked coalescent where each element of $[n]$ additionally has a flag indicating its color: white or blue. Movements and mergers of the blocks of the colored coalescent follow the same dynamics as those of the classical seed bank coalescent. Additionally, if a block activates, each individual inside this block gets the color blue. In other cases the colors remain unchanged.

As in [1], we start with all individuals colored white, so blue only appears after a reactivation event, and we also use the notation $\underline{\mathrm{N}}_n(t)$ (resp. $\underline{\mathrm{M}}_n(t)$) for the number of white plants (resp. white seeds) at time t, starting with n (white) plants and zero seeds.

The notation for the reaching times of $\underline{\mathrm{N}}_n$ are $\underline{\tau}_n^n = 0$ and, for $i \in [n-1]$, $\underline{\tau}_n^i = \inf\{t > 0 : \underline{\mathrm{N}}_n(t) = i\}$.

Note that, on the event $\{\tau_n^i < \theta_n\}$, we have $\underline{\tau}_n^i = \tau_n^i$ almost surely (a.s.), and in general the following stochastic bound holds:

$$\underline{\tau}_n^{i-1} - \underline{\tau}_n^i \leq_{st} \tau_n^{i-1} - \tau_n^i. \tag{3.4}$$

This model is of particular use to prove that the number of seeds that "survive" up to moment θ_n is of order $\log n$. More precisely, as in [1], consider the independent Bernoulli random variables $B_n^i = \mathbf{1}_{\{\text{deactivation at } \underline{\tau}_n^i\}}$ for $i \in [n-1]$, with respective parameter

$$\mathbb{P}\big(B_n^i = 1\big) = \frac{c_1(i+1)}{\binom{i+1}{2} + c_1(i+1)} = \frac{2c_1}{i+2c_1}, \tag{3.5}$$

independently of the number of seeds in the system. It is clear that, almost surely for any $t \geq 0$, $M_n(t) \leq \sum_{i=1}^{n-1} B_n^i$. This and Bienaymé–Chebyshev's inequality lead to the following straightforward result.

Lemma 3.1. *For any $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{t\geq 0} M_n(t) > 2c_1(1+\varepsilon)\log n\right) \leq \frac{1}{2c_1\varepsilon^2 \log n}. \tag{3.6}$$

In particular, for any $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\big(M_n(\theta_n) \leq 2c_1(1+\varepsilon)\log n\big) = 1$.

From now on, denote the upper bound $m_n := \lfloor 2c_1(1+\varepsilon)\log n \rfloor$ for $\varepsilon > 0$. We will use this notation in the following proofs.

The *bounded* seed bank coalescent is a modification of the original seed bank coalescent, where only m seeds can be accumulated in the bank. Thus, when the bank is full, a deactivating lineage disappears instead of moving to the bank. In our case, we start with n plants and m seeds (the bank is full from the beginning).

Denote by $\bar{N}_{n,m}(t)$ (resp. $\bar{M}_{n,m}(t)$) the number of plants (resp. seeds) at time t in the bounded coalescent starting with n plants and m seeds. The block-counting process of the bounded coalescent with parameters $c_1, c_2 > 0$ has the following transition rates for $i \leq n$ and $j \leq m$:

$$\big(\bar{N}_{n,m}(t), \bar{M}_{n,m}(t)\big) \text{ jumps from } (i,j) \text{ to } \begin{cases} (i-1, j) & \text{at rate } \binom{i}{2} + c_1 i \mathbf{1}_{\{j=m\}}, \\ (i-1, j+1) & \text{at rate } c_1 i \mathbf{1}_{\{j<m\}}, \\ (i+1, j-1) & \text{at rate } c_2 j. \end{cases}$$

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

By coupling the seed bank coalescent with its bounded version, we obtain a lower bound for θ_n and an upper bound for $N_n(\theta_n)$.

Lemma 3.2. *Recall T and Z from Proposition 3.1. We have*

$$\lim_{n\to\infty} \mathbb{P}\big(\theta_n \log n \le t\big) \le \mathbb{P}(T \le t), \tag{3.7}$$

$$\lim_{n\to\infty} \mathbb{P}\big(N_n(\theta_n) > z\log n\big) \le \mathbb{P}(Z > z)\,. \tag{3.8}$$

Proof. On the event $\big\{M_n(\theta_n) \le m_n\big\}$, which occurs asymptotically with probability 1 by Lemma 3.1, the variable θ_n is bounded from below, stochastically, by the random variable $\bar\theta_{n,m_n}$ defined by $\bar\theta_{n,m_n} = \inf\big\{t \ge 0 : \bar M_{n,m_n}(t-) > \bar M_{n,m_n}(t)\big\}$ and having exponential distribution with parameter $c_2 m_n$. Then, for $t > 0$,

$$\begin{aligned}\mathbb{P}\big(\theta_n \log n \le t\big) &= \mathbb{P}\big(\theta_n \log n \le t,\, M_n(\theta_n) \le m_n\big) + o(1)\\ &\le \mathbb{P}\big(\bar\theta_{n,m_n} \log n \le t\big) + o(1)\\ &= 1 - \exp\left\{-t\frac{c_2\lfloor 2c_1(1+\varepsilon)\log n\rfloor}{\log n}\right\} + o(1).\end{aligned}$$

So, for any $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\big(\theta_n \log n \le t\big) \le \mathbb{P}(T \le t(1+\varepsilon))$. This gives (3.7).

To prove (3.8), observe that, on the event $\big\{M_n(\theta_n) \le m_n\big\}$, the variable $N_n(\theta_n)$ is bounded from above, stochastically, by the random variable $\bar N_{n,m_n}\big(\bar\theta_{n,m_n}\big)$. So,

$$\mathbb{P}\big(N_n(\theta_n) > z\log n\big) \le \mathbb{P}\big(\bar N_{n,m_n}\big(\bar\theta_{n,m_n}\big) > z\log n\big) + \mathbb{P}\big(M_n(\theta_n) > m_n\big). \tag{3.9}$$

Let us study the asymptotic of $\bar N_{n,m_n}\big(\bar\theta_{n,m_n}\big)$. With similar arguments to Proposition 2.1, we have

$$\begin{aligned}\mathbb{P}\big(\bar N_{n,m_n}\big(\bar\theta_{n,m_n}\big) \le z\log n\big) &= \prod_{i=\lfloor z\log n\rfloor+1}^{n} \frac{\binom{i}{2} + c_1 i}{\binom{i}{2} + c_1 i + c_2 m_n}\\ &= \exp\left\{-\sum_{i=\lfloor z\log n\rfloor+1}^{n} \log\left(1 + \frac{2c_2 m_n}{i(i-1+2c_1)}\right)\right\}\\ &\sim \exp\left\{-2c_2 m_n \sum_{i=\lfloor z\log n\rfloor+1}^{n} \frac{1}{i^2}\right\}.\end{aligned}$$

By a Riemann sum argument, we know that

$$\lim_{n\to\infty} m_n \sum_{i=\lfloor z\log n\rfloor+1}^{n} \frac{1}{i^2} = 2c_1(1+\varepsilon)\int_z^\infty \frac{1}{x^2}\,\mathrm{d}x = \frac{2c_1(1+\varepsilon)}{z}. \tag{3.10}$$

Since $\mathbb{P}(Z \le z) = \exp\{-4c_1c_2/z\}$, we obtain, by taking the limits in (3.9), that $\lim_{n\to\infty} \mathbb{P}\big(N_n(\theta_n) > z\log n\big) \le \mathbb{P}(Z > z/(1+\varepsilon))$, which implies (3.8). □

The bounded seed bank coalescent is also useful to bound from above the random variable $N_n(t)$, for any $t \ge 0$. Let $(K_n(t))_{t\ge0}$ stand for the block-counting process of the Kingman

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

coalescent starting with n lineages. Let $(\chi_i(t))_{i\geq 1}$ be a sequence of i.i.d. Bernoulli variables of parameter $1-\exp(-c_2 t)$. Those variables are more easily understood as $\chi_i(t)=\mathbf{1}_{\{e_i<c_2 t\}}$, where the e_i are i.i.d. standard exponential variables. It is easy to see that, on the event $\{\sup_{t\geq 0} M_n(t)\leq m\}$, stochastically,

$$N_n(t)\leq K_n(t)+\sum_{i=1}^{m}\chi_i(t). \tag{3.11}$$

This follows because $K_n(t)$ bounds the number of blocks that have not been deactivated before time t, and $\sum_{i=1}^{m}\chi_i(t)$ bounds the number of blocks that have already reactivated. Both processes are independent.

We now prove a useful lemma thanks to the two couplings introduced previously. To simplify the notations here and in the followibg, denote $\tau_n^{\lfloor(\log n)^a\rfloor}$ by $\tau_n^{(a)}$, for any $a>0$.

Lemma 3.3. *For $a>b\geq 0$ such that $a+b>1$, $\lim_{n\to\infty}\mathbb{P}\Big(\tau_n^{(a)}\leq(\log n)^{-b}\Big)=1$.*

Proof. Within this proof we denote $m_n:=\lceil 2c_1(1+\varepsilon)\log n\rceil$. Let $E_n=\{\sup_t M_n(t)\leq m_n\}$. We start by observing that $\mathbb{P}\Big(\tau_n^{(a)}>(\log n)^{-b}\Big)=\mathbb{P}\Big(\tau_n^{(a)}>(\log n)^{-b},E_n\Big)+\mathbb{P}\Big(\tau_n^{(a)}>(\log n)^{-b},E_n^{\mathrm{c}}\Big)$. From (3.6), we get

$$\mathbb{P}\big(E_n^{\mathrm{c}}\big)\leq\frac{1}{2c_1\varepsilon^2\log n},$$

so it just remains to control the probability on the event E_n. Recall $(K_n(t))_{t\geq 0}$ and $(\chi_i(t))_{i\geq 1}$ from (3.11). Let $\omega_{n,a}=\inf\{t>0: K_n(t)=\lfloor\frac{1}{2}(\log n)^a\rfloor\}$. Observe that

$$\begin{aligned}\{\tau_n^{(a)}>t,E_n\}&=\{N_n(t)>(\log n)^a,E_n\}\\&\subset\left\{K_n(t)+\sum_{i=1}^{m_n}\chi_i(t)>(\log n)^a\right\}\\&\subset\left\{K_n(t)>\frac{1}{2}(\log n)^a\right\}\cup\left\{\sum_{i=1}^{m_n}\chi_i(t)>\frac{1}{2}(\log n)^a\right\}\\&=\{\omega_{n,a}>t\}\cup\left\{\sum_{i=1}^{m_n}\chi_i(t)>\frac{1}{2}(\log n)^a\right\}.\end{aligned}$$

Taking $t=(\log n)^{-b}$, we obtain $\mathbb{P}\Big(\tau_n^{(a)}>(\log n)^{-b},E_n\Big)\leq\mathbb{P}\big(\omega_{n,a}>(\log n)^{-b}\big)+\mathbb{P}\Big(\sum_{i=1}^{m_n}\chi_i\big((\log n)^{-b}\big)>\frac{1}{2}(\log n)^a\Big)$.

Observe that $\omega_{n,a}$ is the sum of independent exponential random variables with parameter $\binom{i}{2}$ for $\lfloor\frac{1}{2}(\log n)^a\rfloor+1\leq i\leq n$. Thus,

$$\mathbb{E}[\omega_{n,a}]=\sum_{i=\lfloor\frac{1}{2}(\log n)^a\rfloor+1}^{n}\left[\frac{2}{(i-1)}-\frac{2}{i}\right]\leq 2\Big\lfloor\tfrac{1}{2}(\log n)^a\Big\rfloor.$$

So, Markov's inequality for $\omega_{n,a}$ gives $\mathbb{P}\big(\omega_{n,a}>(\log n)^{-b}\big)\leq C(\log n)^{b-a}$ for some constant $C>0$, which converges to 0 whenever $b<a$. On the other hand, Markov's

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

inequality applied to a binomial random variable with parameters $\lfloor 2c_1(1+\varepsilon)\log n\rfloor$ and $1-\exp(-c_2(\log n)^{-b})$ (whose expectation is of order $(\log n)^{1-b}$) leads to

$$\mathbb{P}\left(\sum_{i=1}^{m_n}\chi_i\big((\log n)^{-b}\big) > \frac{1}{2}(\log n)^a\right) \le C(\log n)^{1-b-a}.$$

This quantity converges to 0 when $a+b>1$. □

We now provide the lower bound for $M_n(\theta_n)$. This result, combined with Lemma 3.1 provides the convergence (3.2) in Proposition 3.1.

Lemma 3.4. *For any $\varepsilon>0$ and $a>1$,*

$$\lim_{n\to\infty}\mathbb{P}\big(M_n\big(\tau_n^{(a)}\big) > 2c_1(1-\varepsilon)\log n\big) = 1, \tag{3.12}$$

which implies that

$$\lim_{n\to\infty}\mathbb{P}\big(M_n\big(\theta_n\big) > 2c_1(1-\varepsilon)\log n\big) = 1. \tag{3.13}$$

Proof. Let us first note that (3.8) implies that $\lim_{n\to\infty}\mathbb{P}\big(N_n(\theta_n) < (\log n)^a\big) = 1$, which, thanks to the monotonicity of $(N_n(t))_{t\ge 0}$ until time θ_n, is equivalent to $\lim_{n\to\infty}\mathbb{P}\big(\theta_n > \tau_n^{(a)}\big) = 1$. Due to the monotonicity of $(M_n(t))_{t\ge 0}$ until time θ_n, (3.12) implies (3.13).

Now, on the event $\big\{\theta_n > \tau_n^{(a)}\big\}$, we have $M_n\big(\tau_n^{(a)}\big) = \sum_{i=\lfloor(\log n)^a\rfloor}^{n-1} B_n^i$ almost surely, where the B_n^i are the Bernoulli random variables introduced in (3.5). So,

$$\begin{aligned}\mathbb{P}\big(M_n\big(\tau_n^{(a)}\big) < 2c_1(1-\varepsilon)\log n\big) &= \mathbb{P}\big(M_n\big(\tau_n^{(a)}\big) < 2c_1(1-\varepsilon)\log n, \theta_n > \tau_n^{(a)}\big) + o(1)\\ &\le \mathbb{P}\left(\sum_{i=\lfloor(\log n)^a\rfloor}^{n-1} B_n^i < 2c_1(1-\varepsilon)\log n\right) + o(1).\end{aligned}$$

The latter converges to 0 thanks to the Bienaymé–Chebyshev inequality. □

We are now able to end the overview of the system at time θ_n. The following result, combined with Lemma 3.2, provides the convergences (3.1) and (3.3) in Proposition 3.1.

Lemma 3.5. *Recall T and Z from Proposition 3.1. We have $\lim_{n\to\infty}\mathbb{P}(N_n(\theta_n)\le z\log n) \le \mathbb{P}(Z\le z)$, which implies that*

$$\lim_{n\to\infty}\mathbb{P}\big(\theta_n\log n > t\big) \le \mathbb{P}(T>t). \tag{3.14}$$

Proof. Fix $\varepsilon>0$ and define $\widehat{m}_n := \lfloor 2c_1(1-\varepsilon)\log n\rfloor$. Also, denote $\hat{\tau}_n := \tau_n^{\lfloor z\log n\rfloor}$. First, observe that $\mathbb{P}\big(N_n\big(\theta_n\big)\le z\log n\big) = \mathbb{P}\big(\theta_n \ge \hat{\tau}_n\big)$, so it is enough to prove that

$$\lim_{n\to\infty}\mathbb{P}\big(\theta_n \ge \hat{\tau}_n\big) \le \mathbb{P}(Z\le z). \tag{3.15}$$

For any $t\ge 0$, define $X(t)$ to be the number of reactivations until time t. Let $\mathcal{E}_i$ be an exponential random variable with parameter $c_2 i$, that can be understood as the minimum of i independent exponential random variables with parameter c_2. Then, for any $a>1$,

$$\begin{aligned}\mathbb{P}\big(\theta_n \ge \hat{\tau}_n\big) &= \mathbb{P}(X\big(\hat{\tau}_n\big)=0) = \mathbb{P}\big(X\big(\hat{\tau}_n\big) - X\big(\tau_n^{(a)}\big) = 0, X\big(\tau_n^{(a)}\big)=0\big)\\ &\le \mathbb{P}\big(X\big(\hat{\tau}_n\big) - X\big(\tau_n^{(a)}\big) = 0 \mid X\big(\tau_n^{(a)}\big)=0\big)\\ &\le \mathbb{P}\Big(\mathcal{E}_{M_n\left(\tau_n^{(a)}\right)} > \hat{\tau}_n - \tau_n^{(a)}\Big).\end{aligned}$$

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

The latter inequality follows by observing that if there are no activations in the time interval $\left[\tau_n^{(a)}, \hat{\tau}_n\right]$, then none of the $M_n\left(\tau_n^{(a)}\right)$ seeds present at time $\tau_n^{(a)}$ have activated. Hence,

$$\begin{aligned}\mathbb{P}\left(\theta_n \geq \hat{\tau}_n\right) &\leq \mathbb{E}\left[\mathrm{e}^{-c_2\left(\hat{\tau}_n-\tau_n^{(a)}\right)M_n\left(\tau_n^{(a)}\right)}\right] \\ &= \mathbb{E}\left[\mathrm{e}^{-c_2\left(\hat{\tau}_n-\tau_n^{(a)}\right)M_n\left(\tau_n^{(a)}\right)}\mathbf{1}_{\left\{M_n\left(\tau_n^{(a)}\right)>\widehat{m}_n\right\}}\right] \\ &\quad + \mathbb{E}\left[\mathrm{e}^{-c_2\left(\hat{\tau}_n-\tau_n^{(a)}\right)M_n\left(\tau_n^{(a)}\right)}\mathbf{1}_{\left\{M_n\left(\tau_n^{(a)}\right)\leq\widehat{m}_n\right\}}\right] \\ &\leq \mathbb{E}\left[\mathrm{e}^{-c_2\widehat{m}_n\left(\hat{\tau}_n-\tau_n^{(a)}\right)}\right] + \mathbb{P}\left(M_n\left(\tau_n^{(a)}\right)\leq\widehat{m}_n\right).\end{aligned}$$

So, by denoting for simplicity $n_z = \lfloor z \log n \rfloor$ and $n_a = \lfloor (\log n)^a \rfloor$, and by (3.4), we obtain

$$\mathbb{P}\left(\theta_n \geq \hat{\tau}_n\right) \leq \mathbb{E}\left[\exp\left\{-c_2\widehat{m}_n \sum_{i=n_z+1}^{n_a}\left(\underline{\tau}_n^{i-1} - \underline{\tau}_n^{i}\right)\right\}\right] + \mathbb{P}\left(M_n\left(\tau_n^{(a)}\right)\leq\widehat{m}_n\right). \tag{3.16}$$

Since the variables $\underline{\tau}_n^{i-1} - \underline{\tau}_n^{i}$ are independent and exponentially distributed, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left\{-c_2\widehat{m}_n \sum_{i=n_z+1}^{n_a}\left(\underline{\tau}_n^{i-1} - \underline{\tau}_n^{i}\right)\right\}\right] &= \prod_{i=n_z+1}^{n_a}\frac{\binom{i}{2}+c_1 i}{\binom{i}{2}+c_1 i + c_2\widehat{m}_n} \\ &= \exp\left\{-\sum_{i=n_z+1}^{n_a}\log\left(1+\frac{2c_2\widehat{m}_n}{i(i-1+2c_1)}\right)\right\}.\end{aligned}$$

Now we can use equivalences:

$$\mathbb{E}\left[\exp\left\{-c_2\widehat{m}_n \sum_{i=n_z+1}^{n_a}\left(\underline{\tau}_n^{i-1} - \underline{\tau}_n^{i}\right)\right\}\right] \sim \exp\left\{-\sum_{i=n_z+1}^{n_a}\frac{2c_2\widehat{m}_n}{i^2}\right\}.$$

A similar limit to that given in (3.10) implies that

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}\left[\exp\left\{-c_2\widehat{m}_n \sum_{i=n_z+1}^{n_a}\left(\underline{\tau}_n^{i-1} - \underline{\tau}_n^{i}\right)\right\}\right] &= \exp\left\{-\frac{4c_1c_2(1-\varepsilon)}{z}\right\} \\ &= \mathbb{P}\left(Z \leq \frac{z}{(1-\varepsilon)}\right).\end{aligned} \tag{3.17}$$

Plugging (3.17) and (3.12) into (3.16), and observing that the result is true for any $\varepsilon > 0$, we get (3.15).

A very similar path is followed to obtain (3.14). For some $t > 0$ let $t_n = t(\log n)^{-1}$, and for some $b > 1$ let $s_n = (\log n)^{-b}$. As before, we get

$$\begin{aligned}\mathbb{P}\left(\theta_n \log n > t\right) = \mathbb{P}(\theta_n > t_n) &= \mathbb{P}(X(t_n) = 0) \\ &\leq \mathrm{e}^{-c_2\widehat{m}_n(t_n-s_n)} + \mathbb{P}\left(M_n(s_n)\leq\widehat{m}_n\right).\end{aligned}$$

The first term converges to $\mathbb{P}(T > t(1-\varepsilon))$ and the second to 0. To get the latter, first use (3.7) to see that $\lim_{n\to\infty}\mathbb{P}\left(\theta_n > s_n\right) = 1$. Then, just choose $a > b$ such that Lemma 3.3 holds, and use (3.12). Since the result is true for any $\varepsilon > 0$, we get (3.14). □

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

4. Branch lengths

In this section we study the total branch length L_n of the seed bank coalescent starting with n plants and no seeds, as defined in (1.4), and prove Theorem 1.1 by combining the upcoming Theorems 4.1 and 4.2.

4.1. The active length

Consider the active length defined in (1.5). We prove that this variable has the same (first-order) asymptotics as the total length of the Kingman coalescent.

Theorem 4.1. *Consider the seed bank coalescent starting with n plants and no seeds. Then,* $\lim_{n\to\infty}\{A_n/\log n\} = 2$ *in probability.*

Proof. Recall the notation $\tau_n^{(a)} := \tau_n^{\lfloor(\log n)^a\rfloor}$ and consider some $a \in \left(\frac{1}{2}, 1\right)$. Divide A_n into three parts: $A_n^1 = \int_0^{\theta_n} N_n(t)\,\mathrm{d}t$, $A_n^2 = \int_{\theta_n}^{\tau_n^{(a)}} N_n(t)\,\mathrm{d}t$, and $A_n^3 = \int_{\tau_n^{(a)}}^{\sigma_n} N_n(t)\,\mathrm{d}t$. Here we will work on the event $\left\{\theta_n \le \tau_n^{(a)}\right\}$. The proof for the complementary event follows easily using the same steps. The result is obtained from (4.1), (4.2), and (4.3) in the following.

We first prove that

$$\lim_{n\to\infty} \frac{A_n^1}{\log n} = 2 \quad \text{in probability.} \tag{4.1}$$

Observe that, between times 0 and θ_n, only coalescence or deactivation events occur. This implies that we can rewrite A_n^1 as $A_n^1 = \sum_{i=N_n(\theta_n)+1}^{n} iE_i$, where, given $\left(M_n\left(\tau_n^i\right)\right)_{i=1}^n$, the E_i are independent exponential random variables with respective parameters $\binom{i}{2} + c_1 i + c_2 M_n\left(\tau_n^i\right)$. Indeed, this parameter is that of the minimum of three exponential random variables: the first for coalescence, the second for deactivation, and the third for activation.

Let $h_n = \sum_{i=1}^{n-1}\{2/(i+2c_1)\}$. By proving that $\mathbb{E}\left[|A_n^1 - h_n|\right] = o(\log n)$, we get the desired result. Observe that the variable A_n^1 is stochastically bounded by the length of a Kingman coalescent with freezing (see [18], [6], and [7, Section 1.3]), i.e. $H_n = \sum_{i=2}^{n} iV_i$, where the V_i, as in Section 2, are independent exponential random variables with respective parameters $\binom{i}{2} + c_1 i$. This is true because the seeds 'accelerate' the jump times. To be precise, consider the following coupling. Let $V_i = \min\left\{E_i^{(c)}, E_i^{(d)}\right\}$, where $E_i^{(c)}$ is exponential with parameter $\binom{i}{2}$ and $E_i^{(d)}$ is exponential with parameter $c_1 i$. Now let $E_{i,m}^{(a)}$ be exponential with parameter $c_2 m$. Construct a process $\left(\tilde{N}_n(t), \tilde{M}_n(t)\right)_{t\ge 0}$, equal in distribution to $\left(N_n(t), M_n(t)\right)_{t\ge 0}$ up to time θ_n, recursively, using these exponential random variables:

$$\left(\tilde{N}_n(t), \tilde{M}_n(t)\right) \text{ jumps from } (i,m) \text{ to } \begin{cases} (i-1, m) & \text{if } \min\left\{E_i^{(c)}, E_i^{(d)}, E_{i,m}^{(a)}\right\} = E_i^{(c)}, \\ (i-1, m+1) & \text{if } \min\left\{E_i^{(c)}, E_i^{(d)}, E_{i,m}^{(a)}\right\} = E_i^{(d)}, \\ (0,0) & \text{otherwise.} \end{cases}$$

Here, (0, 0) represents a cemetery state. Note that in distribution $\left(\tilde{N}_n(t), \tilde{M}_n(t)\right) = \left(N_n(t), M_n(t)\right)\mathbf{1}_{\{\theta_n < t\}}$. Thus, by writing $\left(\tilde{\tau}_n^i\right)_{i=1}^n$ for the successive jump times of the new process and $\tilde{r}_n = \sup\left\{i \ge 1 : \min\left\{E_i^{(c)}, E_i^{(d)}, E_{i,\tilde{M}_n\left(\tilde{\tau}_n^i\right)}^{(a)}\right\} = E_{i,\tilde{M}_n\left(\tilde{\tau}_n^i\right)}^{(a)}\right\}$, we obtain $A_n^1 = \sum_{i=\tilde{r}_n+1}^{n} iV_i \le \sum_{i=2}^{n} iV_i = H_n$, where the first equality is in distribution and the others stand almost surely. The first equality is true because, although the V_i are variables with the 'wrong' parameter, they are

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

not independent of $\tilde{r}_n$, and this dependence 'accelerates' these exponential random variables. Hence, $\mathbb{E}\big[\big|A_n^1 - h_n\big|\big] \le \mathbb{E}\big[H_n - A_n^1\big] + \mathbb{E}\big[\big|H_n - h_n\big|\big]$. The second term is bounded thanks to the L^1-convergence of sums of independent exponential variables. For the first term,

$$\begin{aligned}\mathbb{E}\big[H_n - A_n^1\big] &= \mathbb{E}\Big[H_n - \mathbb{E}\Big[A_n^1 | N_n(\theta_n), \big(M_n\big(\tau_n^i\big)\big)_{i\ge 1}\Big]\Big] \\ &= h_n - \mathbb{E}\left[\sum_{i=N_n(\theta_n)+1}^{n} \frac{2}{i - 1 + 2c_1 + \frac{2c_2 M_n\left(\tau_n^i\right)}{i}}\right] \\ &\le h_n - \mathbb{E}\left[\sum_{i=N_n(\theta_n)+1}^{n} \frac{2}{i - 1 + 2c_1 + \frac{2c_2 \sup_t M_n(t)}{i}}\right].\end{aligned}$$

Then, denote $a_n := \lfloor (\log n)^{1+\varepsilon_1} \rfloor$ for some $\varepsilon_1 > 0$, and recall the notation m_n from Section 3. Now, set the event $E_n = \big\{\sup_t M_n(t) \le m_n, N_n(\theta_n) \le a_n\big\}$. We obtain

$$\begin{aligned}\mathbb{E}\big[H_n - A_n^1\big] &\le h_n - \mathbb{E}\left[\mathbf{1}_{E_n} \sum_{i=N_n(\theta_n)+1}^{n} \frac{2}{i - 1 + 2c_1 + \frac{2c_2 \sup_t M_n(t)}{i}}\right] \\ &\le h_n - \mathbb{P}(E_n) \sum_{i=a_n+1}^{n} \frac{2}{i - 1 + 2c_1 + \frac{2c_2 m_n}{i}} \\ &\le h_n - \mathbb{P}(E_n) \sum_{i=a_n+1}^{n} \frac{2}{i - 1 + 2c_1 + \frac{2c_2 m_n}{a_n+1}}.\end{aligned}$$

Since $m_n/(a_n+1) \le C(\log n)^{-\varepsilon_1}$ for some constant C, and $\mathbb{P}(E_n)$ converges to 1 (thanks to Proposition 3.1), we get

$$\mathbb{E}\big[H_n - A_n^1\big] = o(\log n).$$

The L^1-convergence is thus obtained. This implies (4.1).

We now prove that

$$\lim_{n\to\infty} \frac{A_n^2}{\log n} = 0 \quad \text{in probability.} \tag{4.2}$$

It is clear that, almost surely, $A_n^2 \le \tau_n^{(a)}\big(N_n(\theta_n) + M_n(\theta_n)\big)$. Combining Proposition 3.1 and Lemma 3.3 (choosing $b < a$), we obtain the result.

Finally, we prove that

$$\lim_{n\to\infty} \frac{A_n^3}{\log n} = 0 \quad \text{in probability.} \tag{4.3}$$

To this end, denote $U_0 = N_n\big(\tau_n^{(a)}\big) = \lfloor (\log n)^a \rfloor$ (by definition), $V_0 = M_n\big(\tau_n^{(a)}\big)$ (which, by Lemma 3.1, is stochastically bounded by $2c_1(1+\varepsilon)\log n$), and, for any $k \ge 1$, U_k (resp. V_k) as the number of plants (resp. seeds) at the kth event after time $\tau_n^{(a)}$. Each event can be a coalescence, an activation, or a deactivation. Note that the increments of U_k and V_k are in $\{-1, 1\}$. Let S_n be the number of jump times during the interval $\big(\tau_n^{(a)}, \sigma_n\big]$, i.e. $S_n = \inf\{k \ge 1 : U_k + V_k = 1\}$. With these notations, the active branch length on this time interval can be written as $A_n^3 =$

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

$\sum_{k=0}^{S_n-1} U_k E_k$ where, conditional on U_k and V_k, the E_k are independent exponential random variables with respective parameters $\binom{U_k}{2} + c_1 U_k + c_2 V_k$. So, we have

$$\mathbb{E}\left[A_n^3\right] = \mathbb{E}\left[\sum_{k=0}^{S_n-1} \frac{U_k}{\binom{U_k}{2} + c_1 U_k + c_2 V_k}\right].$$

Now define $D_n := |\{k \geq 0 : U_{k+1} - U_k = -1, V_{k+1} - V_k = 1\}|$ as the number of deactivations during this time interval, and observe that

$$\mathbb{E}[D_n] = \mathbb{E}\left[\sum_{k=0}^{S_n-1} \frac{c_1 U_k}{\binom{U_k}{2} + c_1 U_k + c_2 V_k}\right].$$

This implies that $\mathbb{E}\left[A_n^3\right] = (1/c_1)\mathbb{E}[D_n]$. So, it is enough to study the expectation of D_n. We decompose this as $D_n = \sum_{i=2}^{N_n\left(\tau_n^{(a)}\right)+M_n\left(\tau_n^{(a)}\right)} D_n^i$, where D_n^i is the number of deactivations occurring while the total number of lineages equals i, i.e. $D_n^i := |\{k \geq 0 : U_{k+1} - U_k = -1, V_{k+1} - V_k = 1, U_k + V_k = i\}|$. We will bound $\mathbb{E}[D_n]$ thanks to the next model from [1, Definition 4.9].

Let $\left(\widehat{N}_n(t), \widehat{M}_n(t)\right)_{t\geq 0}$ have the same transitions as $\left(N_n(t), M_n(t)\right)_{t\geq 0}$ whenever $\widehat{N}_n(t) \geq \sqrt{\widehat{N}_n(t) + \widehat{M}_n(t)}$. If not, coalescence events are not permitted. For any $i \geq 2$, by [1, Lemma 4.10], $\mathbb{E}\left[D_n^i\right] \leq \mathbb{E}\left[\widehat{D}_n^i\right]$, where $\widehat{D}_n^i$ stands for the number of deactivations in this model while $\widehat{N}_n(t) + \widehat{M}_n(t) = i$. In what follows we will give an idea of why $\mathbb{E}\left[\widehat{D}_n^i\right] = O\left(i^{-1/2}\right)$, implying that $\mathbb{E}[D_n] = O\left((\log n)^{1/2}\right)$, and hence proving (4.3).

The details of the proof, which are unfortunately quite tedious, can be found in the proofs of [1, Lemmas 4.10 and 4.11]. In the following, suppose that $c_1 = c_2 = 1$ for the sake of simplicity.

Fix $i \geq 2$. The variables $\widehat{D}_n^i$ tend to take higher values when coalescences are not permitted; we focus on this case. Thus, suppose that at time t, $\widehat{N}_n(t) + \widehat{M}_n(t)$ reaches i, with $\widehat{N}_n(t-) = \lfloor\sqrt{i}\rfloor + 1 \geq \sqrt{i+1}$. This means that $\widehat{N}_n(t) = \lfloor\sqrt{i}\rfloor \leq \sqrt{i}$. Reactivations are then needed to allow a new coalescence. Conditional on this configuration, the probability that $\widehat{D}_n^i$ equals 0 is equivalent to

$$\frac{i - \lfloor\sqrt{i}\rfloor}{i} \times \frac{\binom{\lfloor\sqrt{i}\rfloor}{2}}{\binom{\lfloor\sqrt{i}\rfloor}{2} + \lfloor\sqrt{i}\rfloor} \sim 1 - \frac{3}{\sqrt{i}} =: p_i.$$

This corresponds approximately to the probability of one reactivation, followed by one coalescence, before one deactivation. So, we have the almost sure bound $\widehat{D}_n^i \leq \sum_{j=0}^{G^i-1} \Delta_j$, where G^i is a geometric random variable with parameter p_i, and the Δ_j give the number of deactivations between each visit of the state $\lfloor\sqrt{i}\rfloor$. The time when coalescence is not allowed is stochastically bounded from above by the time that a random walk that goes up one unit at rate $i - \sqrt{i}$ (rate at of a reactivation) and down at rate $\sqrt{i}$ (rate of a deactivation), started at zero, spends below level $\sqrt{i}$. The random walk has ballistic speed of order i. In particular, it reaches the level $\sqrt{i}$ after $\sqrt{i}/i = 1/\sqrt{i}$ units of time on average. During the period in which coalescence events are not allowed there are always less than $\sqrt{i}$ plants, each of which deactivates at rate $c_1(=1)$. Then, we conclude that, for any j, $\mathbb{E}\left[\Delta_j\right] \leq \frac{1}{\sqrt{i}} \cdot \sqrt{i} = 1$. This uniform bound implies that $\mathbb{E}\left[\widehat{D}_n^i\right] \leq \mathbb{E}[G^i - 1]\mathbb{E}[\Delta_1] = O\left(\frac{1}{\sqrt{i}}\right)$, since $\mathbb{E}[G^i - 1] \sim \frac{3}{\sqrt{i}}$. □

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

4.2. The inactive length

Consider the inactive length defined in (1.5).

Theorem 4.2. *Consider the seed bank coalescent starting with n plants and no seeds. Then,*

$$\lim_{n\to\infty} \frac{I_n}{\log n} = \frac{2c_1}{c_2} \quad \textit{in probability.}$$

Proof. Divide I_n in two parts, $I_n^1 = \int_0^{\theta_n} M_n(t)\,\mathrm{d}t$ and $I_n^2 = \int_{\theta_n}^{\sigma_n} M_n(t)\,\mathrm{d}t$. It is easy to prove that $I_n^1/\log n$ converges to 0 in probability by observing that, almost surely, $I_n^1 \le M_n(\theta_n)\cdot\theta_n$, and using Proposition 3.1.

To study I_n^2, we approximate it by the accumulated time for the $M_n(\theta_n)$ seeds to activate, namely $\tilde{I}_n^2 = \sum_{k=1}^{M_n(\theta_n)} e_k/c_2$, where the e_k are i.i.d. standard exponential random variables. The asymptotics of this random variable are easily obtained by proceeding in two steps, similar to the proof of Proposition 2.2. First, by Proposition 3.1, we have $M_n(\theta_n)/\log n \to 2c_1$ in probability. Second, by denoting $G_n(t) = \sum_{k=1}^{\lfloor t\log n\rfloor} e_k/c_2$, we get

$$\lim_{n\to\infty} \frac{G_n(t)}{\log n} = \frac{t}{c_2}$$

in L^2. By observing that $\tilde{I}_n^2 = G_n(M_n(\theta_n)/\log n)$, we obtain the desired result,

$$\lim_{n\to\infty} \frac{\tilde{I}_n^2}{\log n} = \frac{2c_1}{c_2} \quad \text{in probability.}$$

Finally, the difference between I_n^2 and $\tilde{I}_n^2$ can be bounded by $I_{N_n(\theta_n)} + I_{M_n(\theta_n)}$. Indeed, the variable $I_{N_n(\theta_n)}$ bounds the inactive length resulting from the plants present at time θ_n, and the variable $I_{M_n(\theta_n)}$ bounds the inactive length resulting from the seeds present at time θ_n that activate and deactivate again. Its expectation is clearly of order $\log\log n$. This can be seen by repeating the earlier arguments of this proof. □

5. Sampling formula

Consider the seed bank coalescent at time θ_n and go back through the active part of the genealogical tree until time zero when there are n active lineages and zero inactive lineages. During this period of time we observe $n - N_n(\theta_n)$ events divided into two types: branching inside one lineage (corresponding to a coalescence) and the appearance of a new lineage (corresponding to a deactivation). When there are k lineages, the probability that a branching event occurs is

$$\frac{\binom{k+1}{2}}{\binom{k+1}{2} + c_1(k+1)} = \frac{k}{k+2c_1},$$

whereas the probability that a new lineage appears is $2c_1/(k+2c_1)$. This observation leads to a connection with the classical Hoppe's urn and the Chinese restaurant process (with parameter $2c_1$), which are the key tools to prove Ewens' sampling formula for the law of the allele frequency spectrum in the neutral model, see [7, Section 1.3]. However, in our case, the initial configuration is made of a random number $N_n(\theta_n)$ of tables (old lineages) with one client in each. By applying results of [18], we can obtain a conditional sampling formula corresponding to observing a certain configuration of lineages that passed through the seed bank and lineages that did not deactivate (until time θ_n).

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

Now, let $k \le n$ be a positive integer; we define the sets $A(k,n) = \{a_i, b_i \ge 0, i \in [n] : \sum_{i=1}^n a_i = k \text{ and } \sum_{i=1}^n i(a_i + b_i) = n\}$ and $\bar{A}(k,n) = \{a_i \ge 0, i \in [n] : \sum_{i=1}^n a_i = k \text{ and } \sum_{i=1}^n i a_i \le n\}$. From [18, Eq.(3.3.2)], we obtain the next theorem.

Theorem 5.1. *Let O_i be the number of 'old' blocks of size i (i.e. active blocks of size i at time θ_n) and let R_i be the number of 'recent' blocks of size i (i.e. inactive blocks of size i at time θ_n). Then*

$$\mathbb{P}(O_1 = a_1, \dots, O_n = a_n, R_1 = b_1, \dots, R_n = b_n \mid N_n(\theta_n)) \overset{\text{a.s.}}{=} \frac{(n - N_n(\theta_n))! N_n(\theta_n)!}{(N_n(\theta_n) + 2c_1)_{(n - N_n(\theta_n))}} \prod_{i=1}^n \frac{1}{a_i!} \prod_{j=1}^n \frac{1}{b_j!} \left(\frac{2c_1}{j}\right)^{b_j}, \tag{5.1}$$

with $(a_i, b_i)_{i \in [n]} \in A(N_n(\theta_n), n)$.

Here, the notation $x_{(n)}$ stands for the ascending factorial, i.e. $x_{(n)} = x(x+1)\dots(x+n-1)$.

Remark 5.1. From the latter result and Proposition 3.1, we can obtain an approximate unconditioned sampling formula for large n,

$$\begin{aligned}
&\mathbb{P}(O_1 = a_1, \dots, O_n = a_n, R_1 = b_1, \dots, R_n = b_n) \\
&= \int_0^\infty \mathbb{P}\big(O_1 = a_1, \dots, O_n = a_n, R_1 = b_1, \dots, R_n = b_n \mid N_n(\theta_n) = \lfloor z \log n \rfloor\big) \\
&\qquad \times \mathbb{P}(N_n(\theta_n) = \lfloor z \log n \rfloor)\, \mathrm{d}z \\
&\sim \prod_{i=1}^n \frac{1}{a_i!} \prod_{j=1}^n \frac{1}{b_j!} \left(\frac{2c_1}{j}\right)^{b_j} \\
&\quad \times \int_0^\infty \frac{\Gamma(n - z\log n + 1)\Gamma(z\log n + 1)\Gamma(z\log n + 2c_1)}{\Gamma(n + 2c_1)} \cdot \frac{4c_1c_2}{z^2} e^{-\frac{4c_1c_2}{z}} dz,
\end{aligned}$$

which does not depend on the non-observable variable $N_n(\theta_n)$. Observe that our sampling formula does not make a statement on allele frequencies, but on block frequencies of active and inactive blocks at time θ_n.

The variables O_i and R_i can be inferred if we are capable of deciding if a present individual has visited the seed bank or not. This seems hard, since deactivation can be treated similarly to mutations from the mathematical point of view, but, as opposed to mutations, they do not leave tractable evidence. Furthermore, our result presents a snapshot of the partition at a random time, not at a deterministic one. For these reasons it seems too optimistic to believe that this study provides a possible method of estimating the parameters of the seed bank model.

From (5.1), we obtain the probability-generating function of the old and recent blocks.

Corollary 5.1. *Let $O_1, \dots, O_n, R_1, \dots, R_n$ be random variables with joint density given by (5.1). Then, their (conditional) probability-generating function is*

$$\mathbb{E}\left[\prod_{i=1}^n t_i^{O_i} \prod_{j=1}^n s_j^{R_j} \mid N_n(\theta_n)\right] = \frac{(n - N_n(\theta_n))! N_n(\theta_n)!}{(N_n(\theta_n) + 2c_1)_{(n - N_n(\theta_n))}} \times \sum_{a_1, \dots, a_n, b_1, \dots, b_n \in A(N_n(\theta_n), n)} \prod_{i=1}^n \frac{(t_i)^{a_i}}{a_i!} \prod_{j=1}^n \frac{1}{b_j!} \left(\frac{2c_1 s_j}{j}\right)^{b_j}. \tag{5.2}$$

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

Following the idea of [18] we use two artificial variables, $u \in (-1, 1)$ and $v \in (-1, 1)$. They will help us to rewrite (5.2) in a simpler way. First, observe that, for $(a_i, b_i) \in A(k, n)$,

$$\prod_{i=1}^{n} \left(uv^i\right)^{a_i} \prod_{j=1}^{n} \left(v^j\right)^{b_j} = u^{\sum_{i=1}^{n} a_i} v^{\sum_{i=1}^{n} i(a_i+b_i)} = u^k v^n.$$

Now let $c_{k,n}$ be the multiplying coefficient of $u^k v^n$ in $\exp\left\{\sum_{i=1}^{n} uv^i t_i + \sum_{j=1}^{\infty} \frac{2c_1}{j} s_j v^j\right\}$. We can rewrite (5.2) as

$$\mathbb{E}\left[\prod_{i=1}^{n} t_i^{O_i} \prod_{j=1}^{n} s_j^{R_j} | N_n(\theta_n)\right] = \frac{(n - N_n(\theta_n))! N_n(\theta_n)!}{(N_n(\theta_n) + 2c_1)_{(n-N_n(\theta_n))}} c_{N_n(\theta_n),n}. \tag{5.3}$$

From this relation, we obtain the probability-generating function of the lineages that have not gone through the seed bank at time θ_n.

Corollary 5.2. *Let O_i be the number of 'old' blocks of size i (i.e. active blocks of size i at time θ_n). Then, the joint probability-generating function of $O_1, O_2, \ldots, O_n$ is*

$$\mathbb{E}\left[\prod_{i=1}^{n} t_i^{O_i} \mid N_n(\theta_n)\right] = \sum_{a_1,\ldots,a_n \in \bar{A}(N_n(\theta_n),n)} \frac{N_n(\theta_n)!}{a_1! a_2! \cdots a_n!} t_1^{a_1} t_2^{a_2} \cdots t_n^{a_n} \frac{\binom{2c_1+n-z-1}{n-z}}{\binom{2c_1+n-1}{n-N_n\theta_n}}, \tag{5.4}$$

where $z = \sum_{i=1}^{n} ia_i$.

Proof. First, we will write explicitly the term $c_{k,n}$ when $s_j = 1$ for all j. Observe that

$$\begin{aligned}\exp\left\{\sum_{i=1}^{n} uv^i t_i + \sum_{j=1}^{\infty} \frac{2c_1}{j} v^j\right\} &= (1-v)^{-2c_1} \exp\left\{u \sum_{i=1}^{n} v^i t_i\right\} \\ &= (1-v)^{-2c_1} \sum_{k=0}^{\infty} \frac{\left[u \sum_{i=1}^{n} v^i t_i\right]^k}{k!}.\end{aligned}$$

This implies that the coefficient of u^k in the latter expression is

$$\frac{\left[\sum_{i=1}^{n} \left(v^i t_i\right)\right]^k}{k!} (1-v)^{-2c_1} = \frac{\left[\sum_{i=1}^{n} \left(v^i t_i\right)\right]^k}{k!} \left(\sum_{j=0}^{\infty} \binom{2c_1 + j - 1}{j} v^j\right).$$

We now need to find the coefficient of v^n in the latter expression. First, observe that

$$\left[\sum_{i=1}^{n} \left(v^i t_i\right)\right]^k = \sum_{a_1+\cdots+a_n=k} \frac{k!}{a_1! a_2! \cdots a_n!} t_1^{a_1} t_2^{a_2} \cdots t_n^{a_n} v^z,$$

where $z = \sum_{i=1}^{n} ia_i$. For $z \le n$, the coefficient of v^{n-z} in the expression

$$\left(\sum_{j=0}^{\infty} \binom{2c_1 + j - 1}{j} v^j\right)$$

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

is $\binom{2c_1+n-z-1}{n-z}$. So,

$$c_{k,n} = \frac{1}{k!} \sum_{a_1,\dots,a_n \in \bar{A}(k,n)} \frac{k!}{a_1!a_2!\cdots a_n!} t_1^{a_1} t_2^{a_2} \cdots t_n^{a_n} \binom{2c_1+n-z-1}{n-z}.$$

Thus, replacing $c_{N_n(\theta_n),n}$ and $s_j = 1$ for all j in (5.3), we have the result. □

From the previous corollary we obtain the joint distribution of the lineages which have not gone through the seed bank at time θ_n:

$$\mathbb{P}\big(O_1 = a_1, \dots, O_n = a_n \mid N_n(\theta_n)\big) \overset{\text{a.s.}}{=} \frac{N_n(\theta_n)!}{a_1!a_2!\cdots a_n!} \frac{\binom{2c_1+n-z-1}{n-z}}{\binom{2c_1+n-1}{n-N_n(\theta_n)}}$$

when $a_1, \dots, a_n \in \bar{A}(N_n(\theta_n), n)$.

Now, by taking $t_i = t^i$ and $s_j = 1$ for all $i, j \in [n]$ in (5.3), and finding the corresponding coefficient $c_{N_n(\theta_n),n}$, we obtain the conditional probability-generating function of the number of lineages at time zero that have not been through the seed bank until time θ_n,

$$\mathbb{E}\Big[t^{\sum_{i=1}^n iO_i} \mid N_n(\theta_n)\Big] = \sum_{z=N_n(\theta_n)}^{n} t^z \frac{\binom{2c_1+n-z-1}{n-z}\binom{z-1}{z-N_n(\theta_n)}}{\binom{2c_1+n-1}{n-N_n(\theta_n)}}.$$

Finally, from (5.3), by taking $t_i = 1$ for all $i \in [n]$, and from (5.4), we can find the conditional expectations of O_j and R_j for all $j = 1, 2, \dots, n - N_n(\theta_n)$:

$$\mathbb{E}\big[O_j \mid N_n(\theta_n)\big] = N_n(\theta_n) \frac{\binom{2c_1+n-j-1}{n-j-N_n(\theta_n)+1}}{\binom{2c_1+n-1}{n-N_n(\theta_n)}},$$

$$\mathbb{E}\big[R_j \mid N_n(\theta_n)\big] = \frac{2c_1}{j} \frac{\binom{2c_1+n-j-1}{n-j-N_n(\theta_n)}}{\binom{2c_1+n-1}{n-N_n(\theta_n)}}.$$

Funding information

This project was partially supported by CoNaCyT grant FC-2016-1946 and UNAM-PAPIIT grant IA103820.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] BLATH, J., GONZÁLEZ-CASANOVA, A., KURT, N. AND WILKE-BERENGUER, M. (2016). A new coalescent for seed bank models. *Ann. Appl. Prob.* **26**, 857–891.

[2] DELMAS, J. F., DHERSIN, J. S. AND SIRI-JÉGOUSSE, A. (2008). Asymptotic results on the length of coalescent trees. *Ann. Appl. Prob.* **18**, 997–1025.

[3] DHERSIN, J. S., FREUND, F., SIRI-JÉGOUSSE, A. AND YUAN, L. (2013). On the length of an external branch in the Beta-coalescent. *Stoch. Proc. Appl.* **123**, 1691–1715.

[4] DIEHL, C. S. AND KERSTING, G. (2019). Tree lengths for general Λ-coalescents and the asymptotic site frequency spectrum around the Bolthausen–Sznitman coalescent. *Ann. Appl. Prob.* **29**, 2700–2743.

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press

[5] DRMOTA, M., IKSANOV, A., MÖHLE, M. AND RÖSLER, U. (2007). Asymptotic results concerning the total branch length of the Bolthausen–Sznitman coalescent. *Stoch. Proc. Appl*. **117**, 1404–1421.
[6] DONELLY, P. AND TAVARÉ, S. (1986). The ages of alleles and a coalescent. *Adv. Appl. Prob.* **18**, 1–19.
[7] DURRETT, R. (2008). *Probability Models for DNA Sequence Evolution*. Springer, New York.
[8] FREUND, F. AND SIRI-JÉGOUSSE, A. (2021). The minimal observable clade size of exchangeable coalescents. *Braz. J. Prob. Statist.* **35**, 281–292.
[9] FREUND, F. AND SIRI-JÉGOUSSE, A. (2021). The impact of genetic diversity statistics on model selection between coalescents. *Comput. Statist. Data Anal.* **156**, 107055.
[10] GONZÁLEZ-CASANOVA, A. *et al.* (2014). Strong seed bank effects in bacterial evolution. *J. Theor. Biol.* **356**, 62–70.
[11] HOBOLTH, A., SIRI-JÉGOUSSE, A. AND BLADT, M. (2019). Phase-type distributions in population genetics. *Theor. Pop. Biol.* **127**, 16–32.
[12] KAJ, I., KRONE, S. AND LASCOUX, M. (2001). Coalescent theory for seed bank models. *J. Appl. Prob.* **38**, 285–300.
[13] KERSTING, G. (2012). The asymptotic distribution of the length of beta-coalescent trees. *Ann. Appl. Prob.* **22**, 2086–2107.
[14] LAMBERT, A. AND MA, C. (2015). The coalescent in peripatric metapopulations. *J. Appl. Prob.* **52**, 538–557.
[15] LENNON, J. T. AND JONES, S. E. (2011). Microbial seed banks: The ecological and evolutionary implications of dormancy. *Nat. Rev. Microbiol.* **9**, 119.
[16] MAUGHAN, H. (2007). Rates of molecular evolution in bacteria are relatively constant despite spore dormancy. *Evolution* **61**, 280–288.
[17] TELLIER, A. *et al.* (2011). Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proc. Nat. Acad. Sci*. **108**, 17052–17057.
[18] WATTERSON, G. A. (1984). Lines of descent and the coalescent. *Theor. Pop. Biol.* **26**, 77–92.

https://doi.org/10.1017/jpr.2021.79 Published online by Cambridge University Press