A NEW COALESCENT FOR SEED-BANK MODELS

By Jochen Blath[1], Adrián González Casanova[2],
Noemi Kurt[1] and Maite Wilke-Berenguer[3]

Technische Universität Berlin

We identify a new natural coalescent structure, which we call the *seed-bank coalescent*, that describes the gene genealogy of populations under the influence of a strong seed-bank effect, where "dormant forms" of individuals (such as seeds or spores) may jump a significant number of generations before joining the "active" population. Mathematically, our seed-bank coalescent appears as scaling limit in a Wright–Fisher model with geometric seed-bank age structure if the average time of seed dormancy scales with the order of the total population size N. This extends earlier results of Kaj, Krone and Lascoux [*J. Appl. Probab.* **38** (2011) 285–300] who show that the genealogy of a Wright–Fisher model in the presence of a "weak" seed-bank effect is given by a suitably time-changed Kingman coalescent. The qualitatively new feature of the seed-bank coalescent is that ancestral lineages are independently blocked at a certain rate from taking part in coalescence events, thus strongly altering the predictions of classical coalescent models. In particular, the seed-bank coalescent "does not come down from infinity," and the time to the most recent common ancestor of a sample of size n grows like $\log \log n$. This is in line with the empirical observation that seed-banks drastically increase genetic variability in a population and indicates how they may serve as a buffer against other evolutionary forces such as genetic drift and selection.

1. Introduction. Seed-banks can play an important role in the population genetics of a species, acting as a buffer against evolutionary forces such as random genetic drift and selection as well as environmental variability (see, e.g., [17, 25, 30] for an overview). Their presence typically leads to significantly increased genetic variability respectively effective population size (see, e.g., [18, 21, 26, 27]) and could thus be considered as an important "evolutionary force." In particular, classical mechanisms such as fixation and extinction of genes become more com-

plex: Genetic types can in principle disappear completely from the *active* population at a certain time while returning later due to the germination of seeds or activation of dormant forms.

Seed-banks and dormant forms are known for many taxa. For example, they have been suggested to play an important role in microbial evolution [10, 17], where certain bacterial endospores can remain viable for (in principle arbitrarily) many generations.

Despite the many empirical studies concerned with seed-bank effects, their mathematical modelling in population genetics is still incomplete. Yet probabilistic models, and in particular the Kingman coalescent (cf. [16]) and its relatives, have proven to be very useful tools to understand basic principles of population genetics and the interaction of evolutionary forces [28]. Hence, it is natural to try to incorporate and investigate seed-banks effects in such a probabilistic modelling framework, which is the aim of this article. Before we present our model, we now briefly review some of the mathematical work to date.

In 2001, Kaj, Krone and Lascoux [15] postulated and studied an extension of the classical Wright–Fisher model from [8, 29] that includes seed-banks effects. In their model, each generation consists of a fixed amount of N individuals. Each individual chooses its parent a random amount of generations in the past and copies its genetic type. Here, the number of generations that separates each parent and offspring is understood as the time that the offspring spends as a seed or dormant form. Formally, a parent is assigned to each individual in generation k by first sampling a random number B, which is assumed to be independent and identically distributed for each individual, and then choosing a parent uniformly among the N individuals in generation $k - B$ (note that the case $B \equiv 1$ is just the classical Wright–Fisher model).

Kaj, Krone and Lascoux then prove that if the seed-bank age distribution μ of B is restricted to finitely many generations $\{1, 2, \ldots, m\}$, where m is independent of N, then the ancestral process induced by the seed-bank model converges, after the usual scaling of time by a factor N, to a time changed (delayed) Kingman coalescent, where the coalescent rates are multiplied by $\beta^2 := 1/\mathbb{E}[B]^2$. An increase of the expected value of the seed-bank age distribution thus further decelerates the coalescent, leading to an increase in the effective population size. However, as observed by [30], since the overall coalescent tree structure is retained, this leaves the relative allele frequencies within a sample unchanged. In this scenario we thus speak of a "weak" seed-bank effect.

More generally, [3] show that a sufficient condition for convergence to the Kingman coalescent (with the same scaling and delay) in this setup is that $E[B] < \infty$ with B independent of N.

A different extension of the model by Kaj, Krone and Lascoux is considered in [30], where the authors combine the seed-bank model of [15] with fluctuations in population size. They point out that substantial germ banks with small germination

rates may buffer or enhance the effect of the demography. This indicates that seed-bank effects affect the interplay of evolutionary forces and can have important consequences.

In this respect, [3] show that *strong* seed-bank effects can lead to a behaviour which is very different from the Kingman coalescent. In particular, if the seed-bank age distribution is "heavy-tailed," say, $\mu(k) = L(k)k^{-\alpha}$, for $k \in \mathbb{N}$, where L is a slowly varying function, then if $\alpha < 1$ the expected time for the most recent common ancestor is infinite, and if $\alpha < 1/2$ two randomly sampled individuals do not have a common ancestor at all with positive probability. Hence, this will not only delay, but actually completely alter the effect of random genetic drift.

It can be argued that such extreme behaviour seems artificial (although, as mentioned above, bacterial endospores may stay viable for essentially unlimited numbers of generations). Instead, one can turn to the case $B = B(N)$ and scale the seed-bank age distribution μ with N in order to understand its interplay with other evolutionary forces on similar scales. For example, in [10] and [2] the authors study a seed-bank model with $\mu = \mu(N) = (1 - \varepsilon)\delta_1 + \varepsilon\delta_{N^\beta}, \beta > 0, \varepsilon \in (0, 1)$. They show that for $\beta < 1/3$ the ancestral process converges, after rescaling the time by the nonclassical factor $N^{1+2\beta}$, to the Kingman coalescent, so that the expected time to the most recent common ancestor is highly elevated in this scenario. However, in particular since the above seed-bank Wright–Fisher model is highly non-Markovian, the results in other parameter regimes, in particular $\beta = 1$, are still elusive.

To sum up, while there are mathematical results in the weak seed-bank regime, it appears as if the "right" scaling regimes for stronger seed-bank models, and the potentially new limiting coalescent structures, have not yet been identified. This is in contrast to many other population genetic models, where the interplay of suitably scaled evolutionary forces (such as mutation, genetic drift, selection and migration) often leads to elegant limiting objects, such as the ancestral selection graph [19], or the structured coalescent [12, 20]. A particular problem is the loss of the Markov property in Wright–Fisher models with long genealogical "jumps."

In this paper, we thus propose a new Markovian Wright–Fisher type seed-bank model that allows for a clear forward and backward scaling limit interpretation. In particular, the forward limit in a bi-allelic setup will consist of a pair of (S)DEs describing the allele frequency process of our model, while the limiting genealogy, linked by a duality result, is given by an apparently new coalescent structure which we call *seed-bank coalescent*. In fact, the seed-bank coalescent can be thought of as a structured coalescent of a two island model in a "weak migration regime," in which however coalescences are *completely blocked* in one island. Despite this simple description, the seed-bank coalescent exhibits qualitatively altered genealogical features, both in comparison to the Kingman coalescent and the structured coalescent. In particular, we prove in Theorem 4.1 that the seed-bank coalescent "does not come down from infinity," and in Theorem 4.6 that the expected time to the most recent common ancestor of an n sample is of asymptotic

order $\log\log n$ as n gets large. Interestingly, this latter scale agrees with the one for the Bolthausen–Sznitman coalescent identified by Goldschmidt and Martin [9].

Summarising, the seed-bank coalescent seems to be an interesting and natural scaling limit for populations in the presence of a "strong" seed-bank effect. In contrast to previous genealogies incorporating (weak) seed-bank effects, it is an entirely new coalescent structure and not a time-change of Kingman's coalescent, capturing the essence of seed-bank effects in many relevant situations. We conjecture that the seed-bank coalescent is *universal* in the sense that it is likely to arise as the genealogy of other population models, such as Moran models with a suitable seed-bank component.

The remainder of this paper is organised as follows.

In Section 2, we discuss the Wright–Fisher model with a seed-bank component that has a geometric age structure, and show that its two bi-allelic frequency processes (for "active" individuals and "seeds") converge to a two-dimensional system of (S)DEs. We derive their *dual* process and employ this duality to compute the fixation probabilities as $t \to \infty$ (in law) of the system.

In Section 3, we define the *seed-bank coalescent* corresponding to the previously derived dual block-counting process and show how it describes the ancestry of the Wright–Fisher geometric seed-bank model.

In Section 4, we prove some interesting properties of the seed-bank coalescent, such as "*not* coming down from infinity" and asymptotic bounds on the expected time to the most recent common ancestor, which show that genealogical properties of a population in the presence of strong seed-banks are altered qualitatively.

Sections 3 and 4 may be read essentially independently of Section 2, thus readers who are interested in the coalescent process only may proceed directly to Section 3.

2. The Wright–Fisher model with geometric seed-bank component.

2.1. *The forward model and its scaling limit.*   Consider a haploid population of fixed size N reproducing in fixed discrete generations $k = 0, 1, \ldots$. Assume that individuals carry a genetic type from some type-space E (we will later pay special attention to the bi-allelic setup, say $E = \{a, A\}$, for the forward model).

Further, assume that the population also sustains a *seed-bank* of constant size $M = M(N)$, which consists of the dormant individuals. For simplicity, we will frequently refer to the N "active" individuals as "plants" and to the M dormant individuals as "seeds."

Given $N, M \in \mathbb{N}$, let $\varepsilon \in [0, 1]$ such that $\varepsilon N \leq M$ and set $\delta := \varepsilon N/M$, and assume for convenience that $\varepsilon N = \delta M$ is a natural number (otherwise replace it by $\lfloor \varepsilon N \rfloor$ everywhere). Let $[N] := \{1, \ldots, N\}$ and $[N]_0 := [N] \cup \{0\}$. The dynamics of our Wright–Fisher model with strong seed-bank component are then as follows:

- The N active individuals (plants) from generation 0 produce $(1 - \varepsilon)N$ active individuals in generation 1 by multinomial sampling with equal weights.

- Additionally, $\delta M = \varepsilon N$ uniformly (without replacement) sampled seeds from the seed-bank of size M in generation 0 "germinate," that is, they turn into exactly one active individual in generation 1 each, and leave the seed-bank.
- The active individuals from generation 0 are thus replaced by these $(1 - \varepsilon)N + \delta M = N$ new active individuals, forming the population of plants in the next generation 1.
- Regarding the seed-bank, the N active individuals from generation 0 produce $\delta M = \varepsilon N$ seeds by multinomial sampling with equal weights, filling the vacant slots of the seeds that were activated.
- The remaining $(1 - \delta)M$ seeds from generation 0 remain inactive and stay in the seed-bank (or, equivalently, produce exactly one offspring each, replacing the parent).
- Throughout reproduction, offspring and seeds copy/respectively maintain the genetic type of the parent.

Thus, in generation 1, we have again N active individuals and M seeds. This probabilistic mechanism is then to be repeated independently to produce generations $k = 2, 3, \ldots$. Note that the offspring distribution of active individuals (both for the number of plants and for the number of seeds) is exchangeable within their respective sub-population. Further, one immediately sees that the time that a given seed stays in the seed-bank before becoming active is geometric with success parameter δ, while the probability a given plant produces a dormant seed is ε.

DEFINITION 2.1 (Wright–Fisher model with geometric seed-bank component). Fix population size $N \in \mathbb{N}$, seed-bank size $M = M(N)$, genetic-type space E and δ, ε as before. Given initial type configurations $\xi_0 \in E^N$ and $\eta_0 \in E^M$, denote by

$$\xi_k := \big(\xi_k(i)\big)_{i \in [N]}, \qquad k \in \mathbb{N},$$

the random genetic-type configuration in E^N of the plants in generation k (obtained from the above mechanism), and denote by

$$\eta_k := \big(\eta_k(j)\big)_{j \in [M]}, \qquad k \in \mathbb{N},$$

correspondingly the genetic-type configuration of the seeds in E^M. We call the discrete-time Markov chain $(\xi_k, \eta_k)_{k \in \mathbb{N}_0}$ with values in $E^N \times E^M$ the *type configuration process* of the *Wright–Fisher model with geometric seed-bank component*. See Figure 1 for a possible realisation of the model.

We now specialise to the bi-allelic case $E = \{a, A\}$ and define the frequency chains of a alleles in the active population and in the seed-bank. Define

$$(2.1) \quad X_k^N := \frac{1}{N} \sum_{i \in [N]} \mathbf{1}_{\{\xi_k(i) = a\}} \quad \text{and} \quad Y_k^M := \frac{1}{M} \sum_{j \in [M]} \mathbf{1}_{\{\eta_k(j) = a\}}, \qquad k \in \mathbb{N}_0.$$
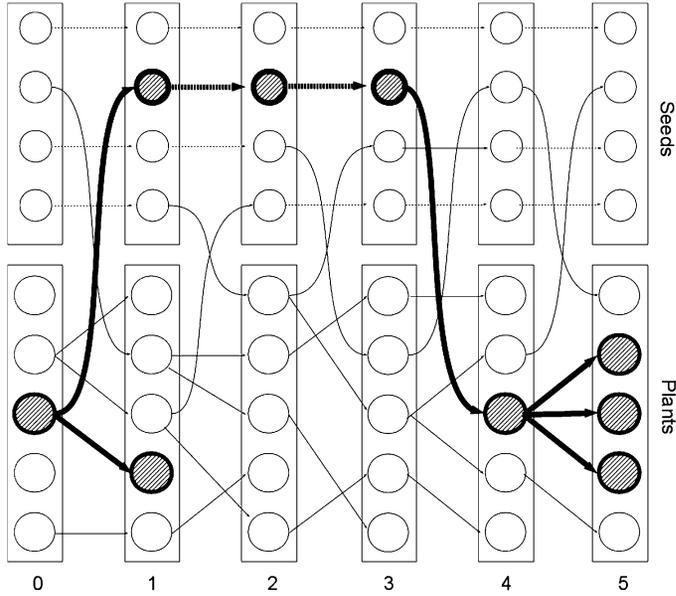
FIG. 1. *A realisation of ancestral relationships in a Wright–Fisher model with geometric seed-bank component. Here, the genetic type of the third plant in generation* 0 (*highlighted in grey*) *is lost after one generation, but returns in generation four via the seed-bank, which acts as a buffer against genetic drift and maintains genetic variability.*

Both are discrete-time Markov chains taking values in

$$I^N = \left\{ 0, \frac{1}{N}, \frac{2}{N}, \ldots, 1 \right\} \subset [0, 1] \quad \text{respectively}$$

$$I^M = \left\{ 0, \frac{1}{M}, \frac{2}{M}, \ldots, 1 \right\} \subset [0, 1].$$

Denote by $\mathbb{P}_{x,y}$ the distribution for which (X^N, Y^M) starts in $(x, y) \in I^N \times I^M$ $\mathbb{P}_{x,y}$-a.s., that is,

$$\mathbb{P}_{x,y}(\cdot) := \mathbb{P}(\cdot | X_0^N = x, Y_0^M = y) \qquad \text{for } (x, y) \in I^N \times I^M$$

(with analogous notation for the expectation, variance, etc.). The corresponding time-homogeneous transition probabilities can now be characterised.

PROPOSITION 2.2. *Let $c := \varepsilon N = \delta M$ and assume $c \in [N]_0$. With the above notation we have for (x, y), respectively, $(\bar{x}, \bar{y}) \in I^N \times I^M$,*

$$p_{x,y} := \mathbb{P}_{x,y}(X_1^N = \bar{x}, Y_1^M = \bar{y})$$

$$= \sum_{i=0}^{c} \mathbb{P}_{x,y}(Z = i)\mathbb{P}_{x,y}(U = \bar{x}N - i)\mathbb{P}_{x,y}(V = \bar{y}M + i),$$
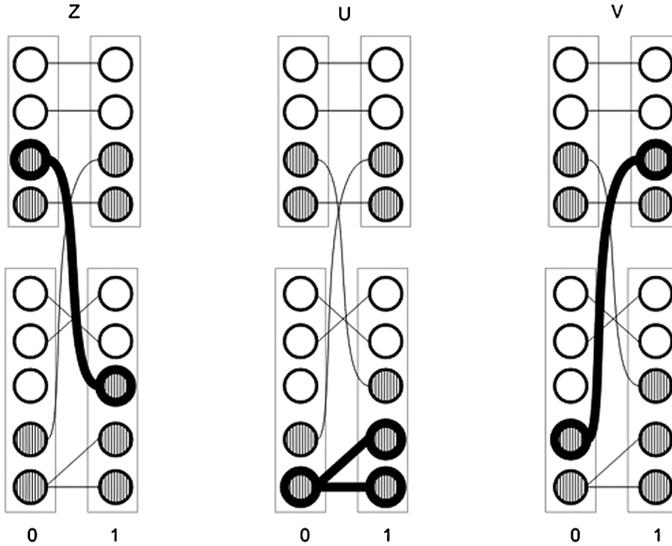
FIG. 2. *In this figure $Z = 1$, $U = 2$ and $V = 1$.*

where Z, U, V are independent under $\mathbb{P}_{x,y}$ with distributions

$$\mathcal{L}_{x,y}(Z) = \mathrm{Hyp}_{M,c,yM}, \qquad \mathcal{L}_{x,y}(U) = \mathrm{Bin}_{N-c,x}, \qquad \mathcal{L}_{x,y}(V) = \mathrm{Bin}_{c,x}.$$

Here, $\mathrm{Hyp}_{M,c,yM}$ *denotes the hypergeometric distribution with parameters M, c, $y \cdot M$ and $\mathrm{Bin}_{c,x}$ is the binomial distribution with parameters c and x.*

REMARK 2.3. The random variables introduced in Proposition 2.2 have a simple interpretation, that is illustrated in Figure 2:

Z is the number of plants in generation 1, that are offspring of a seed of type a in generation 0. This corresponds to the number of seeds of type a that germinate/become active in the next generation (noting that, in contrast to plants, the "offspring" of a germinating seed is always precisely one plant and the seed vanishes).

U is the number of plants in generation 1, that are offspring of plants of type a in generation 0.

V is the number of seeds in generation 1, that are produced by plants of type a in generation 0.

PROOF OF PROPOSITION 2.2. With the interpretation of Z, U and V given in Remark 2.3 their distributions are immediate as described in the Definition 2.1. By construction, we then have $X_1^N = \frac{U+Z}{N}$ and $Y_1^M = y + \frac{V-Z}{M}$ and thus the claim follows. □

In many modelling scenarios in population genetics, parameters describing evolutionary forces such as mutation, selection and recombination are scaled in terms

of the population size N in order to reveal a nontrivial limiting structure (see, e.g., [5] for an overview). In our case, the interesting regime is reached by letting ε, δ (and M) scale with N. More precisely, assume that there exist $c, K \in (0, \infty)$ such that

$$(2.2) \qquad \varepsilon = \varepsilon(N) = \frac{c}{N} \quad \text{and} \quad M = M(N) = \frac{N}{K}.$$

Without loss of generality, $c \in [N]_0$ as $N \to \infty$. Under assumption (2.2), the seed-bank age distribution is geometric with parameter

$$(2.3) \qquad \delta = \delta(N) = \frac{c}{M(N)} = \frac{cK}{N},$$

and c is the number of seeds that become active in each generation, respectively, the number of individuals that move to the seed-bank. The parameter K determines the relative size of the seed-bank with respect to the active population.

PROPOSITION 2.4. *Assume that* (2.2) *holds. Consider test functions* $f \in C^{(3)}([0, 1]^2)$. *For any* $(x, y) \in I^N \times I^M$, *we define the* discrete generator $A^N = A^N_{(\varepsilon, \delta, M)}$ *of the frequency Markov chain* $(X^N_k, Y^M_k)_{k \in \mathbb{N}}$ *by*

$$A^N f(x, y) := N \mathbb{E}_{x, y} [f(X^N_1, Y^M_1) - f(x, y)].$$

Then for all $(x, y) \in [0, 1]^2$,

$$\lim_{N \to \infty} A^N f(x, y) = A f(x, y),$$

where A is defined by

$$A f(x, y) := c(y - x) \frac{\partial f}{\partial x}(x, y) + cK(x - y) \frac{\partial f}{\partial y}(x, y) + \frac{1}{2} x(1 - x) \frac{\partial f^2}{\partial x^2}(x, y).$$

A proof can be found in the Appendix; see Proposition A.1. Since the state space of our frequency chain can be embedded in the compact unit square $[0, 1]^2$, we get tightness and convergence on path-space easily by standard argument (see, e.g., [6], Theorems 4.8.2 and 3.7.8) and can identify the limit of our frequency chains as a pair of the following (S)DEs.

COROLLARY 2.5 (Wright–Fisher diffusion with seed-bank component). *Under the conditions of Proposition* 2.4, *if* $X^N_0 \to x$ *a.s. and* $Y^M_0 \to y$ *a.s., we have that*

$$(X^N_{\lfloor Nt \rfloor}, Y^N_{\lfloor Nt \rfloor})_{t \geq 0} \Rightarrow (X_t, Y_t)_{t \geq 0}$$

on $D_{[0, \infty)}([0, 1]^2)$ *as* $N \to \infty$, *where* $(X_t, Y_t)_{t \geq 0}$ *is a 2-dimensional diffusion solving*

$$(2.4) \qquad \begin{aligned} dX_t &= c(Y_t - X_t) \, dt + \sqrt{X_t(1 - X_t)} \, dB_t, \\ dY_t &= cK(X_t - Y_t) \, dt, \end{aligned}$$

where $(B_t)_{t \geq 0}$ *is standard Brownian motion.*

The proof again follows from standard arguments, cf., for example, [6], where in particular Proposition 2.4 in Chapter 8 shows that the operator A is indeed the generator of a Markov process.

REMARK 2.6. If we abandon the assumption $N = KM$, there are situations in which we can still obtain meaningful scaling limits. If we assume $N/M \to 0$, and we rescale the generator as before by measuring the time in units of size N, we obtain (cf. Proposition A.1)

$$\lim_{N \to \infty} A^N f(x, y) = c(y - x)\frac{\partial f}{\partial x}(x, y) + \frac{1}{2}x(1 - x)\frac{\partial f^2}{\partial x^2}(x, y).$$

This shows that the limiting process is purely one-dimensional, namely the seed-bank frequency Y_t is constantly equal to y, and the process $(X_t)_{t \geq 0}$ is a Wright–Fisher diffusion with migration (with migration rate c and reverting to the mean y). The seed-bank, which in this scaling regime is much larger than the active population, thus acts as a reservoir with constant allele frequency y, with which the plant population interacts.

The case $M/N \to 0$ leads to a simpler limit: If we rescale the generator by measuring the time in units of size M, we obtain

$$\lim_{M \to \infty} A^M f(x, y) = c(y - x)\frac{\partial f}{\partial y}(x, y)$$

and constant frequency $X \equiv x$ in the plant population, which tells us that if the seed-bank is of smaller order than the active population, the genetic configuration of the seed-bank will converge to the genetic configuration of the active population, in a deterministic way.

The above results can be extended to more general genetic types spaces E in a standard way using the theory of measure-valued respectively Fleming–Viot processes. This will be treated elsewhere. Before we investigate some properties of the limiting system, we first derive its *dual* process.

2.2. *The dual of the seed-bank frequency process.* The classical Wright–Fisher diffusion is known to be dual to the block counting process of the Kingman coalescent, and similar duality relations hold for other models in population genetics. Such dual processes are often extremely useful for the analysis of the underlying system, and it is easy to see that our Wright–Fisher diffusion with geometric seed-bank component also has a nice dual.

DEFINITION 2.7. We define the *block-counting process of the seed-bank coalescent* $(N_t, M_t)_{t \geq 0}$ to be the continuous time Markov chain taking values in

$\mathbb{N}_0 \times \mathbb{N}_0$ with transitions

$$
(2.5) \qquad (n, m) \mapsto \begin{cases} (n - 1, m + 1), & \text{at rate } cn, \\ (n + 1, m - 1), & \text{at rate } cKm, \\ (n - 1, m), & \text{at rate } \binom{n}{2}. \end{cases}
$$

Note that the three possible transitions correspond respectively to the drift of the X-component, the drift of the Y-component, and the diffusion part of the system (2.4). This connection is exploited in the following result.

Denote by $\mathbb{P}^{n,m}$ the distribution for which $(N_0, M_0) = (n, m)$ holds $\mathbb{P}^{n,m}$-a.s., and denote the corresponding expected value by $\mathbb{E}^{n,m}$. It is easy to see that, eventually, $N_t + M_t = 1$ (as $t \to \infty$), $\mathbb{P}^{n,m}$-a.s. for all $n, m \in \mathbb{N}_0$. We now show that $(N_t, M_t)_{t \geq 0}$ is the *moment dual* of $(X_t, Y_t)_{t \geq 0}$.

THEOREM 2.8. *For every $(x, y) \in [0, 1]^2$, every $n, m \in \mathbb{N}_0$ and every $t \geq 0$*

$$
(2.6) \qquad \mathbb{E}_{x,y}[X_t^n Y_t^m] = \mathbb{E}^{n,m}[x^{N_t} y^{M_t}].
$$

PROOF.   Let $f(x, y; n, m) := x^n y^m$. Applying for fixed $n, m \in \mathbb{N}_0$ the generator A of $(X_t, Y_t)_{t \geq 0}$ to f acting as a function of x and y gives

$$
\begin{aligned}
Af(x, y) &= c(y - x)\frac{df}{dx} f(x, y) + \frac{1}{2}x(1 - x)\frac{d^2 f}{dx^2} f(x, y) \\
&\quad + cK(x - y)\frac{df}{dy} f(x, y) \\
&= c(y - x)nx^{n-1}y^m + \frac{1}{2}x(1 - x)n(n - 1)x^{n-2}y^m \\
&\quad + cK(x - y)x^n my^{m-1} \\
&= cn(x^{n-1}y^{m+1} - x^n y^m) + \binom{n}{2}(x^{n-1}y^m - x^n y^m) \\
&\quad + cKm(x^{n+1}y^{m-1} - x^n y^m).
\end{aligned}
$$

Note that the right-hand side is precisely the generator of $(N_t, M_t)_{t \geq 0}$ applied to f acting as a function of n and m, for fixed $x, y \in [0, 1]$. Hence, the duality follows from standard arguments; see, for example, [13], Proposition 1.2.   □

2.3. *Long-term behaviour and fixation probabilities.*   The long-term behaviour of our system (2.4) is not obvious. While a classical Wright–Fisher diffusion $(Z_t)_{t \geq 0}$, given by

$$
dZ_t = \sqrt{Z_t(1 - Z_t)} \, dB_t, \qquad Z_0 = z \in [0, 1],
$$

will get absorbed at the boundaries after finite time a.s. (in fact with finite expectation), hitting 1 with probability z, this is more involved for our frequency process in the presence of a strong seed-bank. Nevertheless, one can still compute its fixation probabilities as $t \to \infty$, at least in law. Obviously, $(0, 0)$ and $(1, 1)$ are absorbing states for the system (2.4). They are also the only absorbing states, since absence of drift requires $x = y$, and for the fluctuations to disappear, it is necessary to have $x \in \{0, 1\}$.

PROPOSITION 2.9.   *All mixed moments of* $(X_t, Y_t)_{t \geq 0}$ *solving* (2.4) *converge to the* same *finite limit depending only on* x, y, K. *More precisely, for each fixed* $n, m \in \mathbb{N}$, *we have*

$$(2.7) \qquad \lim_{t \to \infty} \mathbb{E}_{x,y}[X_t^n Y_t^m] = \frac{y + xK}{1 + K}.$$

PROOF.   Let $(N_t, M_t)_{t \geq 0}$ be as in Definition 2.7, started in $(n, m) \in \mathbb{N}_0 \times \mathbb{N}_0$. Let T be the first time at which there is only one particle left in the system $(N_t, M_t)_{t \geq 0}$, that is,

$$T := \inf\{t > 0 : N_t + M_t = 1\}.$$

Note that for any finite initial configuration (n, m), the stopping time T has finite expectation. Now, by Theorem 2.8,

$$\begin{aligned}
\lim_{t \to \infty} \mathbb{E}_{x,y}[X_t^n Y_t^m] &= \lim_{n \to \infty} \mathbb{E}^{n,m}[x^{N_t} y^{M_t}] \\
&= \lim_{t \to \infty} \mathbb{E}^{n,m}[x^{N_t} y^{M_t} | T \leq t] \mathbb{P}^{n,m}(T \leq t) \\
&\quad + \lim_{t \to \infty} \underbrace{\mathbb{E}^{n,m}[x^{N_t} y^{M_t} | T > t]}_{\leq 1} \mathbb{P}^{n,m}(T > t) \\
&= \lim_{t \to \infty} \left(x \mathbb{P}^{n,m}(N_t = 1, T \leq t) + y \mathbb{P}^{n,m}(M_t = 1, T \leq t)\right) \\
&= \lim_{t \to \infty} \left(x \mathbb{P}^{n,m}(N_t = 1) + y \mathbb{P}^{n,m}(M_t = 1)\right) \\
&= \frac{xK}{1 + K} + \frac{y}{1 + K},
\end{aligned}$$

where the last equality holds by convergence to the invariant distribution of a single particle, jumping between the two states "plant" and "seed" at rate c respectively cK, which is given by $(K/(1 + K), 1/(1 + K))$ and independent of the choice of n, m.   □

COROLLARY 2.10 (Fixation in law).   *Given* c, K, (X_t, Y_t) *converges in distribution as* $t \to \infty$ *to a two-dimensional random variable* (X_∞, Y_∞), *whose distribution is given by*

$$(2.8) \qquad \mathcal{L}_{(x,y)}(X_\infty, Y_\infty) = \frac{y + xK}{1 + K} \delta_{(1,1)} + \frac{1 + (1 - x)K - y}{1 + K} \delta_{(0,0)}.$$

Note that this is in line with the classical results for the Wright–Fisher diffusion: As $K \to \infty$ (i.e., the seed-bank becomes small compared to the plant population), the fixation probability of a alleles approaches x. Further, if K becomes small (so that the seed-bank population dominates the plant population), the fixation probability is governed by the initial fraction y of a-alleles in the seed-bank.

PROOF OF COROLLARY 2.10.   It is easy to see that the only two-dimensional distribution on $[0, 1]^2$, for which all moments are constant equal to $\frac{xK+y}{1+K}$, is given by

$$\frac{y + xK}{1 + K} \delta_{(1,1)} + \frac{1 + (1 - x)K - y}{1 + K} \delta_{(0,0)}.$$

Indeed, uniqueness follows from the moment problem, which is uniquely solvable on $[0, 1]^2$. Convergence in law follows from convergence of all moments due to Theorem 3.3.1 in [6] and the Stone–Weierstraß theorem. $\square$

REMARK 2.11 (Almost sure fixation).   Observing that $(KX_t + Y_t)_{t \geq 0}$ is a bounded martingale, and given the shape of the limiting law (2.8), one can also get almost sure convergence of (X_t, Y_t) to (X_∞, Y_∞) as $t \to \infty$. However, as we will see later, fixation will not happen in finite time, since the block-counting process $(N_t, M_t)_{t \geq 0}$, started from an infinite initial state, *does not come down from infinity* (see Section 4), which means that the whole (infinite) population does not have a most-recent common ancestor. Thus, in finite time, initial genetic variability should never be completely lost. We expect that with some extra work, this intuitive reasoning could be made rigorous in an almost sure sense with the help of a "look-down construction," and will be treated in future work. The fact that fixation does not occur in finite time can also be understood from (2.4), where we can compare the seed-component $(Y_t)_{t \geq 0}$ to the solution of the deterministic equation

$$dy_t = -cKy_t \, dt,$$

corresponding to a situation where the drift towards 0 is maximal [or to $dy_t = cK(1 - y_t) \, dt$ where the drift towards 1 is maximal]. Since $(y_t)_{t \geq 0}$ does not reach 0 in finite time if $y_0 > 0$, neither does $(Y_t)_{t \geq 0}$.

3. The seed-bank coalescent.

3.1. *Definition and genealogical interpretation.*   In view of the form of the block counting process, it is now easy to guess the stochastic process describing the limiting gene genealogy of a sample taken from the Wright–Fisher model with seed-bank component. Indeed, for $k \geq 1$, let $\mathcal{P}_k$ be the set of partitions of $[k]$. For $\pi \in \mathcal{P}_k$ let $|\pi|$ be the number of blocks of the partition π. We define the space of *marked* partitions to be

$$\mathcal{P}_k^{\{p,s\}} = \{(\zeta, \vec{u}) | \zeta \in \mathcal{P}_k, \vec{u} \in \{s, p\}^{|\zeta|}\}.$$

This enables us to attach to each partition block a flag which can be either "plant" or "seed" (p or s), so that we can trace whether an ancestral line is currently in the active or dormant part of the population. For example, for $k = 5$, an element π of $\mathcal{P}_k^{\{p,s\}}$ is the marked partition $\pi = \{\{1, 3\}^p \{2\}^s \{4, 5\}^p\}$.

Consider two marked partitions $\pi, \pi' \in \mathcal{P}_k^{\{p,s\}}$, we say $\pi \succ \pi'$ if π' can be constructed by merging exactly 2 blocks of π carrying the p-flag, and the resulting block in π' obtained from the merging both again carries a p-flag. For example,

$$\{\{1, 3\}^p \{2\}^s \{4, 5\}^p\} \succ \{\{1, 3, 4, 5\}^p \{2\}^s\}.$$

We use the notation $\pi \bowtie \pi'$ if π' can be constructed by changing the flag of precisely one block of π, for example,

$$\{\{1, 3\}^p \{2\}^s \{4, 5\}^p\} \bowtie \{\{1, 3\}^s \{2\}^s \{4, 5\}^p\}.$$

DEFINITION 3.1 (The seed-bank k-coalescent).   For $k \geq 2$ and $c, K \in (0, \infty)$, we define the *seed-bank k-coalescent* $(\Pi_t^{(k)})_{t \geq 0}$ with seed-bank intensity c and relative seed-bank size $1/K$ to be the continuous time Markov chain with values in $\mathcal{P}_k^{\{p,s\}}$, characterised by the following transitions:

$$(3.1) \qquad \pi \mapsto \pi' \text{ at rate } \begin{cases} 1, & \text{if } \pi \succ \pi', \\ c, & \text{if } \pi \bowtie \pi' \text{ and one } p \text{ is replaced by one } s, \\ cK, & \text{if } \pi \bowtie \pi' \text{ and one } s \text{ is replaced by one } p. \end{cases}$$

If $c = K = 1$, we speak of the *standard seed-bank k-coalescent*. See Figure 3 for a possible realisation.

Comparing (3.1) to (2.5), it becomes evident that (N_t, M_t) introduced in Definition 2.7 is indeed the block-counting process of the seed-bank coalescent.

DEFINITION 3.2 (The seed-bank coalescent).   We may define the *seed-bank coalescent*, $(\Pi_t)_{t \geq 0} = (\Pi_t^{(\infty)})_{t \geq 0}$ with seed-bank intensity c and relative seed-bank size $1/K$ as the unique Markov process distributed according to the projective limit as k goes to infinity of the laws of the seed-bank k-coalescents (with seed-bank intensity c and relative seed-bank size $1/K$). In analogy to Definition 3.1, we call the case of $c = K = 1$ the *standard seed-bank coalescent*.

REMARK 3.3.   Note that the seed-bank coalescent is a well-defined object. Indeed, for the projective limiting procedure to make sense, we need to show *consistency* and then apply the Kolmogorov extension theorem. This can be roughly sketched as follows. Define the process $(\overrightarrow{\Pi}_t^{(k)})_{t \geq 0}$ as the projection of $(\Pi_t^{(k+1)})_{t \geq 0}$, the $k + 1$ seed-bank coalescent, to the space $\mathcal{P}_k^{\{p,s\}}$. Mergers and flag-flips involving the singleton $\{k + 1\}$ are only visible in $(\Pi_t^{(k+1)})_{t \geq 0}$, but do not af-
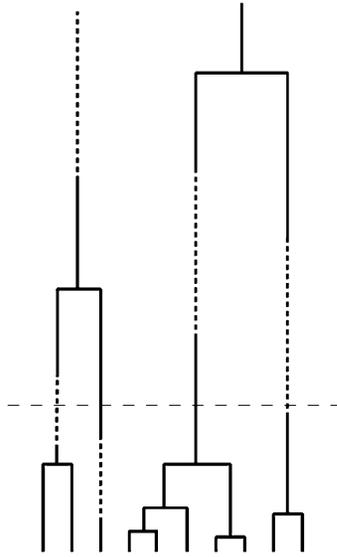
FIG. 3. *A possible realisation of the standard* 10-*seed-bank coalescent. Dotted lines indicate "inactive lineages"* (*carrying an s-flag, which are prohibited from merging*). *At the time marked with the dotted horizontal line the process is in the state* $\{\{1, 2\}^s\{3\}^p\{4, 5, 6, 7, 8\}^p\{9, 10\}^s\}$.

fect $(\overrightarrow{\Pi}_t^{(k)})_{t \geq 0}$. Indeed, by the Markov-property, a change involving the singleton $\{k + 1\}$ does not affect any of the other transitions. Hence, if $\overrightarrow{\Pi}_0^{(k)} = \Pi_0^{(k)}$, then

$$\big(\overrightarrow{\Pi}_t^{(k)}\big)_{t \geq 0} = \big(\Pi_t^{(k)}\big)_{t \geq 0}$$

holds in distribution. By the Kolmogorov extension theorem, the projective limit exists and is unique.

Note that it is obvious that the distribution of the block counting process of the seed-bank coalescent, counting the number of blocks carrying the p and s-flags, respectively, agrees with the distribution the process $(N_t, M_t)_{t \geq 0}$ from Definition 2.7 (with suitable initial conditions).

Further, it is not hard to see that the seed-bank coalescent appears as the limiting genealogy of a sample taken from the Wright–Fisher model with geometric seed-bank component in the same way as the Kingman coalescent describes the limiting genealogy of a sample taken from the classical Wright–Fisher model (here, we merely sketch a proof, which is entirely standard).

Indeed, consider the genealogy of a sample of $k \ll N$ individuals, sampled from present generation 0. We proceed backward in time, keeping track in each generation of the ancestors of the original sample among the active individuals (plants) and among the seeds. To this end, denote by $\Pi_i^{(N,k)} \in \mathcal{P}_k^{\{p,s\}}$ the configuration of the genealogy at generation $-i$, where two individuals belong to the same block

of the partition $\Pi_i^{(N,k)}$ if and only if their ancestral lines have met until generation $-i$, which means that all individuals of a block have exactly one common ancestor in this generation, and the flag s or p indicates whether said ancestor is a plant or a seed in generation $-i$. According to our forward in time population model, there are the following possible transitions from one generation to the previous one of this process:

- One (or several) plants become seeds in the previous generation.
- One (or several) seeds become plants in the previous generations.
- Two (or more) individuals have the same ancestor in the previous generation (which by construction is necessarily a plant), meaning that their ancestral lines merge.
- Any possible combination of these three events.

It turns out that only three of the possible transitions play a role in the limit as $N \to \infty$, whereas the others have a probability that is of smaller order.

PROPOSITION 3.4. *In the setting of Proposition* 2.4, *additionally assume that* $\Pi_0^{(N,k)} = \{\{1\}^p, \ldots, \{k\}^p\}$, $\mathbb{P}$-*a.s. for some fixed* $k \in \mathbb{N}$. *Then for* $\pi, \pi' \in \mathcal{P}_k^{\{p,s\}}$,

$$
(3.2) \quad \begin{aligned}
&\mathbb{P}(\Pi_{i+1}^{(N,k)} = \pi' | \Pi_i^{(N,k)} = \pi) \\
&= \begin{cases}
\dfrac{1}{N} + O(N^{-2}), & \text{if } \pi \succ \pi', \\[2mm]
\dfrac{c}{N} + O(N^{-2}), & \text{if } \pi \bowtie \pi' \text{ and a } p \text{ is replaced by an } s, \\[2mm]
\dfrac{cK}{N} + O(N^{-2}), & \text{if } \pi \bowtie \pi' \text{ and an } s \text{ is replaced by a } p, \\[2mm]
O(N^{-2}), & \text{otherwise}
\end{cases}
\end{aligned}
$$

for all $i \in \mathbb{N}_0$.

PROOF. According to the definition of the forward in time population model, exactly c out of the N plants become seeds, and exactly c out of the $M = N/K$ seeds become plants. Thus, whenever the current state $\Pi_i^{(N,k)}$ of the genealogical process contains at least one p-block, then the probability that a *given* p-block changes flag to s at the next time step is equal to $\frac{c}{N}$. If there is at least one s-block, then the probability that any given s-block changes flag to p is given by $\frac{cK}{N}$, and the probability that a given p-block chooses a *fixed* plant ancestor is equal to $(1 - \frac{c}{N})\frac{1}{N}$ (where $1 - c/N$ is the probability that the ancestor of the block in question is a plant, and $1/N$ is the probability to choose one particular plant among the N).

From this, we conclude that the probability of a coalescence of two given p-blocks in the next step is

$$
\mathbb{P}(\text{two given } p\text{-blocks merge}) = \left(1 - \frac{c}{N}\right)^2 \frac{1}{N}.
$$

Since we start with k blocks, and the blocks move independently, the probability that two or more blocks change flag at the same time is of order at most N^{-2}. Similarly, the probability of any combination of merger or block-flip events other than single blocks flipping or binary mergers is of order N^{-2} or smaller, since the number of possible events (coalescence or change of flag) involving at most k blocks is bounded by a constant depending on k but not on N. $\square$

COROLLARY 3.5. *For any $k \in \mathbb{N}$, under the assumptions of Proposition 2.4, $(\Pi_{\lfloor Nt \rfloor}^{(N,k)})_{t \geq 0}$ converges weakly as $N \to \infty$ to the seed-bank coalescent $(\Pi_t^{(k)})_{t \geq 0}$ started with k plants.*

PROOF. From Proposition 3.4, it is easy to see that the generator of $(\Pi_{\lfloor Nt \rfloor}^{(N,k)})$ converges to the generator of $(\Pi_t^{(k)})$, which is defined via the rates given in (3.1). Then standard results (see Theorem 3.7.8 in [6]) yield weak convergence of the process. $\square$

3.2. *Related coalescent models.* *The structured coalescent.* The seed-bank coalescent is reminiscent of the *structured coalescent* arising from a two-island population model (see, e.g., [11, 12, 20, 24, 29]). Indeed, consider two Wright–Fisher type (sub-)populations of fixed relative size evolving on separate "islands", where individuals (resp., ancestral lineages) may migrate between the two locations with a rate of order of the reciprocal of the total population size (the so-called "weak migration regime"). Since offspring are placed on the same island as their parent, mergers between two ancestral lineages are only allowed if both are currently in the same island. This setup again gives rise to a coalescent process defined on "marked partitions," with the marks indicating the location of the ancestral lines among the two islands. Coalescences are only allowed for lines carrying the same mark at the same time, and marks are switched according to the scaled migration rates. See [28] for an overview.

In our Wright–Fisher model with geometric seed-bank component, we consider a similar "migration" regime between the two sub-populations, in our case called "plants" and "seeds." However, in the resulting seed-bank coalescent, coalescences can only happen while in the plant-population. This asymmetry leads to a behaviour that is qualitatively different to the usual two-island scenario (e.g., with respect to the time to the most recent common ancestor, whose expectation is always finite for the structured coalescent, even if the sample size goes to infinity).

The coalescent with freeze. Another related model is the *coalescent with freeze* (see [4]), where blocks become completely inactive at some rate. This model is different from ours because once a block has become inactive, it cannot be activated again. Hence, it cannot coalesce at all, which clearly leads to a different long-time behaviour. In particular, one will not expect to see a most recent common ancestor in such a coalescent.

4. Properties of the seed-bank coalescent.

4.1. *Coming down from infinity.* The notion of *coming down from infinity* was discussed by Pitman [22] and Schweinsberg [23]. They say that an exchangeable coalescent process *comes down from infinity* if the corresponding block counting process (of an infinite sample) has finitely many blocks immediately after time 0 (i.e., the number of blocks is finite almost surely for each $t > 0$). Further, the coalescent is said to *stay infinite* if the number of blocks is infinite a.s. for all $t \geq 0$. Schweinsberg also gives a necessary and sufficient criterion for so-called "Lambda-coalescents" to come down from infinity. In particular, the Kingman coalescent does come down from infinity. However, note that the seed-bank coalescent does not belong to the class of Lambda-coalescents, so that Schweinsberg's result does not immediately apply. For an overview of the properties of general exchangeable coalescent processes, see, for example, [1].

THEOREM 4.1. *The seed-bank coalescent does* not *come down from infinity. In fact, its block-counting process* $(N_t, M_t)_{t \geq 0}$ *stays infinite for every* $t \geq 0$, $\mathbb{P}$*-a.s. To be precise, for each starting configuration* (n, m) *where* $n + m$ *is* (*countably*) *infinite*,

$$\mathbb{P}\big(\forall t \geq 0 : M_t^{(n,m)} = \infty\big) = 1.$$

The proof of this theorem is based on a coupling with a dominated simplified *coloured seed-bank coalescent* process introduced below. In essence, the coloured seed-bank coalescent behaves like the normal seed-bank coalescent, except we mark the individuals with a colour to indicate whether they have (entered and) left the seed-bank at least once. This will be useful in order to obtain a process where the number of plant-blocks is nonincreasing. We will then prove that even if we consider only those individuals that have never made a transition from seed to plant (but possibly from plant to seed), the corresponding block-counting process will stay infinite. This will be achieved by proving that infinitely many particles enter the seed-bank before any positive time. Since they subsequently leave the seed-bank at a linear rate, this will take an infinite amount of time.

DEFINITION 4.2 (A coloured seed-bank coalescent). In analogy to the construction of the seed-bank coalescent, we first define the set of *coloured*, marked partitions as

$$\mathcal{P}_k^{\{p,s\} \times \{w,b\}} := \big\{(\pi, \vec{u}, \vec{v}) | (\pi, \vec{u}) \in \mathcal{P}_k^{\{p,s\}}, \vec{v} \in \{w, b\}^k\big\}, \qquad k \in \mathbb{N},$$

$$\mathcal{P}^{\{p,s\} \times \{w,b\}} := \big\{(\pi, \vec{u}, \vec{v}) | (\pi, \vec{u}) \in \mathcal{P}^{\{p,s\}}, \vec{v} \in \{w, b\}^{\mathbb{N}}\big\}.$$

It corresponds to the marked partitions introduced earlier, where now each element of $[k]$, respectively, $\mathbb{N}$, has an additional flag indicating its colour: w for *white*

and b for *blue*. We write $\pi \succ \pi'$, if π' can be constructed from π by merging two blocks with a p-flag in π that result into a block with a p-flag in π', while each individual retains its colour. It is important to note that the p- or s-flags are assigned to *blocks*, the colour-flags to *individuals*, that is, elements of $[k]$, respectively, $\mathbb{N}$. We use $\pi \ltimes \pi'$, to denote that π' results from π by changing the flag of a block from p to s and leaving the colours of all individuals unchanged and $\pi \rtimes \pi'$, if π' is obtained from π, by changing the flag of a block from s to p *and colouring all the individuals in this block blue*, that is, setting their individual flags to b. In other words, after leaving the seed-bank, individuals are always coloured blue.

For $k \in \mathbb{N}$ and $c, K \in (0, \infty)$ we now define the *coloured seed-bank k-coalescent with seed-bank intensity c and seed-bank size* $1/K$, denoted by $(\underline{\Pi}_t)_{t \geq 0}$, as the continuous time Markov chain with values in $\mathcal{P}_k^{\{p,s\} \times \{w,b\}}$ and transition rates given by

$$
(4.1) \qquad \pi \mapsto \pi' \text{ at rate } \begin{cases} 1, & \text{if } \pi \succ \pi', \\ c, & \text{if } \pi \ltimes \pi', \\ cK, & \text{if } \pi \rtimes \pi'. \end{cases}
$$

The *coloured seed-bank coalescent with seed-bank intensity c and seed-bank size* $1/K$ is then the unique Markov process on $\mathcal{P}^{\{p,s\} \times \{w,b\}}$ given by the projective limit of the distributions of the k-coloured seed-bank coalescents, as k goes to infinity.

REMARK 4.3.   1. Note that the coloured seed-bank coalescent is well-defined. Since the colour of an individual only depends on its own path and does not depend on the colour of other individuals (not even those that belong to the same block), the consistency of the laws of the k-coloured seed-bank coalescents boils down to the consistency of the seed-bank k-coalescents discussed in Remark 3.3. In much the same way, we then obtain the existence and uniqueness of the coloured seed-bank coalescent from Kolmogorov's Extension theorem.

2. The normal seed-bank (k-)coalescent can be obtained from the coloured seed-bank (k-)coalescent by omitting the flags indicating the colouring of the individuals. However, if we only consider those blocks containing *at least* one white individual, we obtain a coalescent similar to the seed-bank coalescent, where lineages are discarded once they leave the seed-bank.

For $t \geq 0$ define $\underline{N}_t$ to be the number of *white plants* and $\underline{M}_t$ the number of *white seeds* in $\underline{\Pi}_t$. We will use a superscript (n, m) to denote the processes started with n plants and m seeds $\mathbb{P}$-a.s., where $n, m = \infty$ means we start with a countably infinite number of plants, respectively, seeds. We will always start in a configuration were all individual labels are set to w, that is, with only white particles. Note that our construction is such that $(\underline{N}_t)_{t \geq 0}$ is nonincreasing.

PROPOSITION 4.4. *For any $n, m \in \mathbb{N} \cup \{\infty\}$, the processes $(N_t^{(n,m)},$ $M_t^{(n,m)})_{t \geq 0}$ and $(\underline{N}_t^{(n,m)}, \underline{M}_t^{(n,m)})_{t \geq 0}$ can be coupled such that*

$$\mathbb{P}\big(\forall t \geq 0 : N_t^{(n,m)} \geq \underline{N}_t^{(n,m)} \text{ and } M_t^{(n,m)} \geq \underline{M}_t^{(n,m)}\big) = 1.$$

PROOF. This result is immediate if we consider the coupling through the coloured seed-bank coalescent and the remarks in Remark 4.3. □

PROOF OF THEOREM 4.1. Proposition 4.4 implies that it suffices to prove the statement for $(\underline{M}_t)_{t \geq 0}$ instead of $(M_t)_{t \geq 0}$. In addition, we will only have to consider the case of $m = 0$, since starting with more (possibly infinitely many) seeds will only contribute towards our desired result.

For $n \in \mathbb{N} \cup \{\infty\}$, let

$$\tau_j^n := \inf\{t \geq 0 : \underline{N}_t^{(n,0)} = j\}, \qquad 1 \leq j \leq n - 1, j < \infty$$

be the first time that the number of active blocks of an n-sample reaches k. Note that $(\underline{N}_t)_{t \geq 0}$ behaves like the block-counting process of a Kingman coalescent where in addition to the coalescence events, particles may "disappear" at a rate proportional to the number of particles alive. Since the corresponding values for a Kingman coalescent are finite $\mathbb{P}$-a.s., it is easy to see that the τ_j^n are, too. Clearly, for any n, $\tau_j^n - \tau_{j-1}^n$ has an exponential distribution with parameter

$$\lambda_j := \binom{j}{2} + cj.$$

At each time of a transition τ_j^n, we distinguish between two events: *coalescence* and *deactivation* of an active block, where by deactivation we mean a transition of $(\underline{N}_t^n, \underline{M}_t^n)_{t \geq 0}$ of type $(j + 1, l) \mapsto (j, l + 1)$ (for suitable $l \in [n]$), that is, the transition of a plant to a seed.

Then

(4.2) $$\mathbb{P}\big(\text{deactivation at } \tau_{j-1}^n\big) = \frac{cj}{\binom{j}{2} + cj} = \frac{2c}{j + 2c - 1},$$

independently of the number of inactive blocks. Thus,

$$X_j^n := \mathbf{1}_{\{\text{deactivation at } \tau_{j-1}^n\}}, \qquad j = 2, \ldots, n, j < \infty,$$

are independent Bernoulli random variables with respective parameters $2c/(j + 2c - 1)$, $j = 2, \ldots, n$. Note that X_j^n depends on j, but the random variable is independent of the random variable τ_{j-1} due to the memorylessness of the exponential distribution. Now define A_t^n as the (random) number of deactivations up to time $t \geq 0$ that is, for $n \in \mathbb{N} \cup \{\infty\}$,

(4.3) $$A_t^n := \sum_{j=2}^{n} X_j^n \mathbf{1}_{\{\tau_{j-1}^n < t\}}.$$

For $n \in \mathbb{N}$, since $\lambda_j \geq \binom{j}{2}$, it follows from a comparison with the block counting process of the Kingman coalescent, denoted by $(|\tilde{\Pi}_t^n|)_{t \geq 0}$ if started in n blocks, that for all $t \geq 0$,

$$\lim_{n \to \infty} \mathbb{P}(\tau_{\lfloor \log n - 1 \rfloor}^n \leq t) \geq \lim_{n \to \infty} \mathbb{P}(|\tilde{\Pi}_t^n| \leq \lfloor \log n - 1 \rfloor)$$

$$\geq \lim_{n \to \infty} \mathbb{P}(|\tilde{\Pi}_t| \leq \log n - 1) = 1,$$

where the last equality follows from the fact that the Kingman coalescent $(\tilde{\Pi}_t)_{t \geq 0}$ comes down from infinity, cf. [22, 23]. For $t \geq 0$,

$$(4.4) \qquad \mathbb{P}\left(A_t^n \geq \sum_{j=\log n}^n X_j^n \right) \geq \mathbb{P}\left(\mathbf{1}_{\{\tau_{\log n - 1}^n < t\}} \sum_{j=\log n}^n X_j^n \geq \sum_{j=\log n}^n X_j^n \right)$$

$$(4.5) \qquad \geq \mathbb{P}(\tau_{\log n - 1}^n < t)$$

and hence, by (4.4)

$$(4.6) \qquad \lim_{n \to \infty} \mathbb{P}\left(A_t^n \geq \sum_{j=\log n}^n X_j^n \right) = 1.$$

Note that due to (4.2),

$$(4.7) \quad \mathbb{E}\left[\sum_{j=\log n}^n X_j^n \right] = \sum_{j=\log n}^n \frac{2c}{j + 2c - 1} = 2c(\log n - \log \log n) + R(c, n),$$

where $R(c, n)$ converges to a finite value depending on the seed-bank intensity c as $n \to \infty$. Since the X_j^n are independent Bernoulli random variables, we obtain for the variance

$$\mathbb{V}\left[\sum_{j=\log n}^n X_j^n \right] = \sum_{j=\log n}^n \mathbb{V}[X_j^n] = \sum_{j=\log n}^n \frac{2c}{j + 2c - 1}\left(1 - \frac{2c}{j + 2c - 1} \right)$$

$$(4.8) \qquad \qquad \leq 2c \log n \qquad \text{as } n \to \infty.$$

For any $\varepsilon > 0$, we can choose n large enough such that, $\mathbb{E}[\sum_{j=\log n}^n X_k] \geq (2c - \varepsilon) \log n$ holds, which yields

$$\mathbb{P}\left(\sum_{j=\log n}^n X_j^n < c \log n \right) \leq \mathbb{P}\left(\sum_{j=\log n}^n X_j^n - \mathbb{E}\left[\sum_{j=\log n}^n X_j^n \right] < -(c - \varepsilon) \log n \right)$$

$$(4.9) \qquad \qquad \leq \mathbb{P}\left(\left| \sum_{j=\log n}^n X_j^n - \mathbb{E}\left[\sum_{j=\log n}^n X_j^n \right] \right| > (c - \varepsilon) \log n \right)$$

$$\leq \frac{2c}{(c - \varepsilon)^2 \log n},$$

by Chebyshev's inequality. In particular, for any $\kappa \in \mathbb{N}$,

$$\lim_{n \to \infty} \mathbb{P}\left(\sum_{j=\log n}^{n} X_j^n < \kappa \right) = 0,$$

and together with (4.6) we obtain for any $t > 0$

$$(4.10) \qquad \lim_{n \to \infty} \mathbb{P}(A_t^n < \kappa) = 0.$$

Since the $(A_t^n)_{t \geq 0}$ are coupled by construction for any $n \in \mathbb{N} \cup \{\infty\}$, we know in particular that $\mathbb{P}(A_t^\infty < \kappa) \leq \mathbb{P}(A_t^n < \kappa)$, for any $n \in \mathbb{N}, t \geq 0, \kappa \geq 0$ and therefore $\mathbb{P}(A_t^\infty < \kappa) = 0$, which yields

$$(4.11) \qquad \forall t \geq 0 \qquad \mathbb{P}(A_t^\infty = \infty) = 1.$$

Since in addition, $(A_t^\infty)_{t \geq 0}$ is nondecreasing in t, we can even conclude

$$(4.12) \qquad \mathbb{P}(\forall t \geq 0 : A_t^\infty = \infty) = 1.$$

Thus, we have proven that, for any time $t \geq 0$, there have been an infinite amount of movements to the seed-bank $\mathbb{P}$-a.s. Now we are left to show that this also implies the presence of an infinite amount of lineages in the seed-bank, that is, that a sufficiently large proportion is saved from moving back to the plants where it would be "instantaneously" reduced to a finite number by the coalescence mechanism.

Define $\mathcal{B}_t$ to be the blocks of a partition that visited the seed-bank at some point before a fixed time $t \geq 0$ and were visible in the "white" seed-bank coalescent, that is,

$$\mathcal{B}_t := \{ B \subseteq \mathbb{N} | \exists 0 \leq r \leq t : B^{\{s\}} \in \underline{\Pi}_r^{(\infty,0)}$$

$$\text{and contains at least one } white \text{ particle} \}.$$

Since we started our coloured coalescent in $(\infty, 0)$, the cardinality of $\mathcal{B}_t$ is at least equal to A_t^∞ and, therefore, we know $\mathbb{P}(|\mathcal{B}_t| = \infty) = 1$. Since $\mathcal{B}_t$ is countable, we can enumerate its elements as $\mathcal{B}_t = \bigcup_{n \in \mathbb{N}} \{B_t^n\}$ and use this to define the sets $\mathcal{B}_t^n := \{B_t^1, \ldots, B_t^n\}$, for all $n \in \mathbb{N}$. Since $\mathcal{B}_t$ is infinite $\mathbb{P}$-a.s., these $\mathcal{B}_t^n$ exist for any n, $\mathbb{P}$-a.s. Now observe that the following inequalities hold even pathwise by construction:

$$\underline{M}_t^{(\infty,0)} \geq \sum_{B \in \mathcal{B}_t} \mathbf{1}_{\{B^{\{s\}} \in \underline{\Pi}_t^{(\infty,0)}\}} \geq \sum_{B \in \mathcal{B}_t^n} \mathbf{1}_{\{B^{\{s\}} \in \underline{\Pi}_t^{(\infty,0)}\}}$$

and, therefore, the following holds for any $\kappa \in \mathbb{N}$:

$$\mathbb{P}(\underline{M}_t^{(\infty,0)} \leq \kappa) \leq \mathbb{P}\left(\sum_{B \in B_t^n} \mathbf{1}_{\{B^s \in \underline{\Pi}_t^{(\infty,0)}\}} \leq \kappa \right)$$

$$\overset{*}{\leq} \sum_{i=1}^{\kappa} \binom{n}{i} (e^{-ct})^i (1 - e^{-ct})^{n-i} \overset{n \to \infty}{\longrightarrow} 0,$$

which in turn implies $\mathbb{P}(\underline{M}_t^{(\infty,0)} = \infty) = 1$. In * we used that for each of the n blocks in $\mathcal{B}_t^n$ we know $\mathbb{P}(B \in \underline{\Pi}_t^{(\infty,0)}) \geq e^{-ct}$ and they leave the seed-bank independently of each other, which implies that the sum is dominated by a Binomial random variable with parameters n and e^{-ct}.

Since the probability on the left does not depend on n, and the above holds for any $\kappa \in \mathbb{N}$, we obtain $\mathbb{P}(\underline{M}_t^{(\infty,0)} = \infty) = 1$ for all $t > 0$. Note that this also implies $\mathbb{P}(\underline{M}_t^{(\infty,0)} + \underline{N}_t^{(\infty,0)} = \infty) = 1$ for all $t > 0$, from which, through the monotonicity of the sum, we can immediately deduce the stronger statement

$$\mathbb{P}(\forall t > 0 : \underline{M}_t^{(\infty,0)} + \underline{N}_t^{(\infty,0)} = \infty) = 1.$$

On the other hand, we have seen that $\mathbb{P}(\underline{N}_t^{(\infty,0)} < \infty) = 1$, for all $t > 0$, which again using its monotonicity, yields $\mathbb{P}(\forall t > 0 : \underline{N}_t^{(\infty,0)} < \infty) = 1$. Putting these two results together, we obtain $\mathbb{P}(\forall t > 0 : \underline{M}_t^{(\infty,0)} = \infty) = 1$. $\quad\square$

4.2. *Bounds on the time to the most recent common ancestor.* In view of the previous subsection, it is now quite obvious that the seed-bank causes a relevant delay in the time to the most recent common ancestor of finite samples. Throughout this section, we will again use the notation $(N_t^{(n,m)}, M_t^{(n,m)})$ to indicate the initial condition of the block counting process is (n, m).

DEFINITION 4.5. We define the *time to the most recent common ancestor* of a sample of n plants and m seeds, to be

$$T_{\mathrm{MRCA}}[(n, m)] = \inf\{t > 0 : (N_t^{(n,m)}, M_t^{(n,m)}) = (1, 0)\}.$$

Since coalescence only happens in the plants, $T_{\mathrm{MRCA}}[(n, m)] = \inf\{t > 0 : N_t^{(n,m)} + M_t^{(n,m)} = 1\}$. We will mostly be interested in the case where the sample is drawn from plants only, and write $T_{\mathrm{MRCA}}[n] := T_{\mathrm{MRCA}}[(n, 0)]$. The main results of this section are asymptotic logarithmic bounds on the expectation of $T_{\mathrm{MRCA}}[n]$.

THEOREM 4.6. *For all $c, K \in (0, \infty)$, the seed-bank coalescent satisfies*

$$(4.13) \qquad\qquad \mathbb{E}[T_{\mathrm{MRCA}}[n]] \asymp \log\log n.$$

Here, the symbol $\asymp$ denotes weak asymptotic equivalence of sequences, meaning that we have

$$(4.14) \qquad\qquad \liminf_{n\to\infty} \frac{\mathbb{E}[T_{\mathrm{MRCA}}[n]]}{\log\log n} > 0$$

and

$$(4.15) \qquad\qquad \limsup_{n\to\infty} \frac{\mathbb{E}[T_{\mathrm{MRCA}}[n]]}{\log\log n} < \infty.$$

The proof of Theorem 4.6 will be given in Propositions 4.8 and 4.12. The intuition behind this result is the following. The time until a seed gets involved in a coalescence event is much longer than the time it takes for a plant to be involved in a coalescence, since a seed has to become a plant first. Thus, the time to the most recent common ancestor of a sample of n plants is governed by the number of individuals that become seeds before coalescence, and by the time of coalescence of a sample of seeds.

Due to the quadratic coalescence rates, it is clear that the time until the ancestral lines of all sampled plants have either coalesced into one, or have entered the seed-bank at least once, is finite almost surely. The number of lines that enter the seed-bank until that time is a random variable that is asymptotically of order $\log n$, due to similar considerations as in (4.7). Thus, we need to control the time to the most recent common ancestor of a sample of $O(\log n)$ seeds. The linear rate of migration then leads to the second log.

Turning this reasoning into bounds requires some more work, in particular for an upper bound. As in the proof of Theorem 4.1, let $X_k, k = 1, \ldots, n$ denote independent Bernoulli random variables with parameters $2c/(k + 2c - 1)$. Similar to (4.3) define

$$(4.16) \qquad\qquad A^n := \sum_{k=2}^{n} X_k.$$

LEMMA 4.7. *Under our assumptions, for any $\varepsilon > 0$,*

$$\lim_{n \to \infty} \mathbb{P}(A^n \geq (2c + \varepsilon) \log n) = 0$$

and

$$\lim_{n \to \infty} \mathbb{P}(A^n \leq (2c - \varepsilon) \log n) = 0.$$

PROOF. As in the proof of Theorem 4.1 before, we have

$$\mathbb{E}[A^n] = \sum_{k=2}^{n} \frac{2c}{k + 2c - 1} = 2c \log n + R'(c, n),$$

where $R'(c, n)$ converges to a finite value depending on c as $n \to \infty$, and

$$\mathbb{V}(A^n) \sim 2c \log n \qquad \text{as } n \to \infty.$$

Thus, again by Chebyshev's inequality, for sufficiently large n (and recalling that c is our model parameter)

$$\begin{aligned}
\mathbb{P}(A^n \geq (2c + \varepsilon) \log n) &\leq \mathbb{P}(A^n - \mathbb{E}[A^n] \geq \varepsilon \log n) \\
&\leq P(|A^n - \mathbb{E}[A^n]| \geq \varepsilon \log n) \\
&\leq \frac{2c}{\varepsilon^2 \log n}.
\end{aligned}$$

This proves the first claim. The second statement follows similarly, cf. (4.9).   □

Recall the process $(\underline{N}_t, \underline{M}_t)_{t \geq 0}$ from the previous subsection. The coupling of Proposition 4.4 leads to the lower bound in Theorem 4.6.

PROPOSITION 4.8.   *For all $c, K \in (0, \infty)$, the seed-bank coalescent satisfies*

$$(4.17) \qquad \liminf_{n \to \infty} \frac{\mathbb{E}[T_{\mathrm{MRCA}}[n]]}{\log \log n} > 0.$$

PROOF.   The coupling with $(\underline{N}_t, \underline{M}_t)_{t \geq 0}$ yields

$$T_{\mathrm{MRCA}}[n] \geq \underline{T}_{\mathrm{MRCA}}[n],$$

where $\underline{T}_{\mathrm{MRCA}}[n]$ denotes the time until $(\underline{N}_t, \underline{M}_t)$ started at $(n, 0)$ has reached a state with only one block left. By definition, A^n of the previous lemma gives the number of individuals that at some point become seeds in the process $(\underline{N}_t, \underline{M}_t)_{t \geq 0}$. Thus, $\underline{T}_{\mathrm{MRCA}}[n]$ is bounded from below by the time it takes until these A^n seeds migrate to plants (and then disappear). Since the seeds disappear independently of each other, we can bound $\underline{T}_{\mathrm{MRCA}}[n]$ stochastically from below by the extinction time of a pure death process with death rate cK started with A^n individuals. For such a process started at $A^n = l \in \mathbb{N}$ individuals, the expected extinction time as $l \to \infty$ is of order $\log l$. Thus, we have for $\varepsilon > 0$ that there exists $C > 0$ such that

$$\mathbb{E}[T_{\mathrm{MRCA}}[n]] \geq \mathbb{E}[T_{\mathrm{MRCA}}[n]\mathbf{1}_{\{A^n \geq (2c - \varepsilon) \log n\}}]$$
$$\geq C \log \log n \mathbb{P}(A^n \geq (2c - \varepsilon) \log n),$$

and the claim follows from the fact that by Lemma 4.7, $A^n \geq (c - \varepsilon) \log n$ almost surely as $n \to \infty$.   □

To prove the corresponding upper bound, we couple (N_t, M_t) to a functional of another type of coloured process.

DEFINITION 4.9.   Let $(\overline{N}_t, \overline{M}_t)_{t \geq 0}$ be the continuous-time Markov process with state space $E \subset \mathbb{N} \times \mathbb{N}$, characterised by the transition rates:

$$(n, m) \mapsto \begin{cases} (n - 1, m + 1), & \text{at rate } cn, \\ (n + 1, m - 1), & \text{at rate } cKm, \\ (n - 1, m), & \text{at rate } \binom{n}{2} \cdot \mathbf{1}_{\{n \geq \sqrt{n + m}\}}. \end{cases}$$

This means that $(\overline{N}_t, \overline{M}_t)_{t \geq 0}$ has the same transitions as (N_t, M_t), but coalescence is suppressed if there are too few plants relative to the number of seeds. The effect of this choice of rates is that for $(\overline{N}_t, \overline{M}_t)_{t \geq 0}$, if $n \gtrsim \sqrt{m}$, then coalescence happens at a rate which is of the same order as the rate of migration from seed to plant.

LEMMA 4.10. *The processes* $(\overline{N}_t, \overline{M}_t)_{t \geq 0}$ *and* $(N_t, M_t)_{t \geq 0}$ *can be coupled such that*

$$\mathbb{P}\big(\forall t \geq 0 : N_t^{(n,m)} \leq \overline{N}_t^{(n,m)} \text{ and } M_t^{(n,m)} \leq \overline{M}_t^{(n,m)}\big) = 1.$$

PROOF. We construct both processes from the same system of blocks. Start with $n + m$ blocks labelled from $\{1, \ldots, n + m\}$, and with n of them carrying an s-flag, the others a p-flag. Let $S^i, P^i, i = 1, \ldots, n + m$ and $V^{i,j}, i, j = 1, \ldots, n + m, i < j$ be independent Poisson processes, S^i with parameter cK, P^i with parameter c, and $V^{i,j}$ with parameter 1. Moreover, let each block carry a colour flag, blue or white. At the beginning, all blocks are supposed to be blue. The blocks evolve as follows: At an arrival of S^i, if block i carries an s-flag, this flag is changed to p irrespective of the colour and the state of any other block. Similarly, at an arrival of P^i, if block i carries a p-flag, this is changed to an s-flag. At an arrival of $V^{i,j}$, and if blocks i and j both carry a p-flag, one observes the whole system, and proceeds as follows:

(i) If the total number of p-flags in the system is greater or equal to the square root of the total number of blocks, then blocks i and j coalesce, which we encode by saying that the block with the higher label (i or j) is discarded. If the coalescing blocks have the same colour, this colour is kept. Note that here the *blocks* carry the colour, unlike in the coloured process of the previous sections, where the individuals were coloured. If the coalescing blocks have different colours, then the colour after the coalescence is blue.

(ii) If the condition on the number of flags in (i) is not satisfied, then there is no coalescence, but if both blocks were coloured blue, then the block (i or j) with the higher label is coloured white (this can be seen as a "hidden coalescence" in the process where colours are disregarded).

It is then clear by observing the rates that (N_t, M_t) is equal in distribution to the process which counts at any time t the number of blue blocks with p-flags and with s-flags, respectively, and $(\overline{N}_t, \overline{M}_t)$ is obtained by counting the number of p-flags and s-flags of any colour. By construction we obviously have $\overline{N}_t \geq N_t$ and $\overline{M}_t \geq M_t$ for all t. □

Define now

$$\overline{T}_{\text{MRCA}}[m] := \inf\{t \geq 0 : \big(\overline{N}_t^{(0,m)}, \overline{M}_t^{(0,m)}\big) = (1, 0)\}.$$

LEMMA 4.11. *There exists a finite constant C independent of m such that*

$$\overline{T}_{\text{MRCA}}[m] \leq C \log m.$$

PROOF. Define for every $k \in 1, 2, \ldots, m - 1$ the hitting times

$$(4.18) \qquad\qquad H_k := \inf\{t > 0 : \overline{N}_t + \overline{M}_t = k\}.$$

We aim at proving that $\mathbb{E}^{0,m}[H_{m-1}] \leq \frac{C}{\sqrt{m}}$ and $\mathbb{E}^{0,m}[H_{j-1} - H_j] \leq \frac{C}{j-1}$ for $j \leq m - 1$, for some $0 < C < \infty$. Here and throughout the proof, C denotes a generic positive constant (independent of m) which may change from instance to instance. To simplify notation, we will identify $\sqrt{j}$ with $\lceil \sqrt{j} \rceil$, or equivalently assume that all occurring square roots are natural numbers. Moreover, we will only provide the calculations in the case of the standard seed-bank coalescent, that is, $c = K = 1$. The reader is invited to convince herself that the argument can also be carried out in the general case.

We write $\overline{\lambda}_t$ for the total jump rate of the process $(\overline{N}, \overline{M})$ at time t, that is,

$$\overline{\lambda}_t = \binom{\overline{N}_t}{2} 1_{\{\overline{N}_t \geq \sqrt{\overline{N}_t + \overline{M}_t}\}} + \overline{N}_t + \overline{M}_t,$$

and set

$$\overline{\alpha}_t := \frac{\binom{\overline{N}_t}{2} 1_{\{\overline{N}_t \geq \sqrt{\overline{N}_t + \overline{M}_t}\}}}{\overline{\lambda}_t}, \qquad \overline{\beta}_t := \frac{\overline{N}_t}{\overline{\lambda}_t}, \qquad \overline{\gamma}_t := \frac{\overline{M}_t}{\overline{\lambda}_t}$$

for the probabilities that the first jump after time t is a coalescence, a migration from plant to seed or a migration from seed to plant, respectively. Even though all these rates are now random, they are well-defined conditional on the state of the process. The proof will be carried out in three steps.

Step 1: *Bound on the time to reach $\sqrt{m}$ plants.* Let

(4.19) $$D_m := \inf\{t > 0 : \overline{N}_t^{(0,m)} \geq \sqrt{m}\}$$

denote the first time the number of plants is at least $\sqrt{m}$. Due to the restriction in the coalescence rate, the process $(\overline{N}_t^{(0,m)}, \overline{M}_t^{(0,m)})_{t \geq 0}$ has to first reach a state with at least $\sqrt{m}$ plants before being able to coalesce, hence $D_m < H_{m-1}$ a.s. Hence, for any $t \geq 0$, conditional on $t \leq D_m$ we have $\overline{\lambda}_t = m$ and $\overline{N}_t < \sqrt{m}$. Thus, $\overline{M}_t > m - \sqrt{m}$ a.s. and we note that at each jump time of $(\overline{N}_t, \overline{M}_t)$ for $t \leq D_m$

$$\overline{\gamma}_s \geq \frac{m - \sqrt{m}}{m} = 1 - \frac{1}{\sqrt{m}} \qquad \text{a.s. } \forall s \leq t$$

and

$$\overline{\beta}_s \leq \frac{1}{\sqrt{m}} \qquad \text{a.s. } \forall s \leq t.$$

The expected number of jumps of the process $(\overline{N}_t, \overline{M}_t)$ until D_m is therefore bounded from above by the expected time it takes a discrete time asymmetric simple random walk started at 0 with probability $1 - 1/\sqrt{m}$ for an upward jump and $1/\sqrt{m}$ for a downward jump to reach level $\sqrt{m} - 1$. It is a well-known fact (see, e.g., [7], Chapter XIV.3) that this expectation is bounded by $C\sqrt{m}$ for some

$C \in (0, \infty)$. Since the time between each of the jumps of $(\overline{N}_t, \overline{M}_t)$, for $t < D_m$, is exponential with rate $\overline{\lambda}_t = m$, we get

$$(4.20) \qquad \mathbb{E}^{0,m}[D_m] \leq C\sqrt{m} \cdot \frac{1}{m} = \frac{C}{\sqrt{m}}.$$

Step 2: *Bound on the time to the first coalescence after reaching $\sqrt{m}$ plants.* At time $t = D_m$, we have $\overline{\lambda} = \binom{\sqrt{m}}{2} + \sqrt{m} + m - \sqrt{m}$, and thus

$$\overline{\beta}_t = \frac{\sqrt{m}}{\binom{\sqrt{m}}{2} + m} = \frac{2\sqrt{m}}{3m - \sqrt{m}} \leq \frac{C}{\sqrt{m}} \qquad \text{a.s.}$$

and

$$\overline{\alpha}_t = \frac{m - \sqrt{m}}{3m - \sqrt{m}} \geq \frac{1}{3}\left(1 - \frac{1}{\sqrt{m}}\right) \qquad \text{a.s.}$$

Denote by J_m the time of the first jump after time D_m. At J_m there is either a coalescence taking place (thus reaching a state with $m - 1$ individuals and hence in that case $H_{m-1} = J_m$), or a migration. In order to obtain an upper bound on H_{m-1}, as a "worst-case scenario," we can assume that if there is no coalescence at J_m, the process is restarted from state $(0, m)$, and then run again until the next time that there are at least $\sqrt{m}$ plants (hence after J_m, the time until this happens is again equal in distribution to D_m). If we proceed like this, we have that the number of times that the process is restarted is stochastically dominated by a geometric random variable with parameter $\frac{1}{3}(1 - \frac{1}{\sqrt{m}})$, and since

$$\mathbb{E}^{0,m}[J_m - D_m] = \lambda_{D_m}^{-1} = \frac{1}{\binom{\sqrt{m}}{2} + m} \leq \frac{C}{m},$$

we can conclude [using (4.20)] that

$$\mathbb{E}^{0,m}[H_{m-1}] \leq \mathbb{E}^{0,m}[J_m]\frac{3\sqrt{m}}{\sqrt{m} - 1}$$

$$(4.21) \qquad = \left(\mathbb{E}^{0,m}[D_m] + \mathbb{E}^{0,m}[J_m - D_m]\right)\frac{3\sqrt{m}}{\sqrt{m} - 1}$$

$$\leq \frac{C}{\sqrt{m}}.$$

Step 3: *Bound on the time between two coalescences.* Now we want to estimate $\mathbb{E}^{0,m}[H_{j-1} - H_j]$ for $j \leq m - 1$. Obviously, at time H_j-, for $j \leq m - 1$, there are at least $\sqrt{j+1}$ plants, since $\overline{N}_t + \overline{M}_t$ can decrease only through a coalescence. Therefore (keeping in mind our convention that Gauß-brackets are applied if necessary, and hence $\overline{N}_{H_j} \geq \sqrt{j+1} - 1 \geq \sqrt{j} - 1$ holds) we obtain $\overline{N}_{H_j} \geq \sqrt{j} - 1$. Hence, either we have $\overline{N}_{H_j} \geq \sqrt{j}$ and coalescence is possible in the first jump

after H_j, or $\overline{N}_{H_j} = \sqrt{j} - 1$, in which case $\overline{\gamma}_{H_j} \geq \frac{j - \sqrt{j}}{j} = 1 - \frac{1}{\sqrt{j}}$, meaning that if coalescence is not allowed at H_j, with probability at least $1 - \frac{1}{\sqrt{j}}$ it will be possible after the first jump after reaching H_j. Thus, the probability that coalescence is allowed either at the first or the second jump after time H_j is bounded from below by $1 - \frac{1}{\sqrt{j}}$.

Assuming that coalescence is possible at the first or second jump after H_j, denote by L_j the time to either the first jump after H_j if $\overline{N}_{H_j} \geq \sqrt{j}$, or the time of the second jump after H_j otherwise. Then in the same way as before, we see that $\overline{\alpha}_{L_j} \geq 1 - \frac{C}{\sqrt{j}}$. Thus, the probability that H_{j-1} is reached no later than two jumps after H_j is at least $(1 - \frac{C}{\sqrt{j}})^2$. Otherwise, in the case where there was no coalescence at either the first or the second jump after H_j, we can obtain an upper bound on H_{j-1} by restarting the process from state $(0, j)$. The probability that the process is restarted is thus bounded from above by $\frac{C}{\sqrt{j}}$. We know from equation (4.21) that if started in $(0, j)$, there is $\mathbb{E}^{0,j}[H_{j-1}] \leq \frac{C}{\sqrt{j}}$. Noting that $\overline{\lambda}_{H_j} \geq j$, and we need to make at most two jumps, we have that $\mathbb{E}^{0,m}[L_j] \leq 2/j$. Thus, we conclude

$$\mathbb{E}^{0,m}[H_{j-1} - H_j] \leq \mathbb{E}^{0,m}[L_j]\left(1 - \frac{C}{\sqrt{j}}\right)^2 + \frac{C}{\sqrt{j}}\mathbb{E}^{0,j}[H_{j-1}]$$

$$(4.22) \qquad\qquad \leq \frac{2}{j-1}\left(1 - \frac{C}{\sqrt{j}}\right) + \left(\frac{C}{\sqrt{j}}\right)^2$$

$$\leq \frac{C}{j-1}.$$

These three bounds allow us to complete the proof, since when starting $(\overline{N}_t, \overline{M}_t)$ in state $(0, m)$ our calculations show that

$$\mathbb{E}[\overline{T}_{\text{MRCA}}[m]] = \mathbb{E}^{0,m}[H_1] = \mathbb{E}[H_m] + \sum_{j=2}^{m-1}\mathbb{E}[H_{j-1} - H_j]$$

$$(4.23) \qquad\qquad \leq \frac{C}{\sqrt{m}} + C\sum_{j=2}^{m-1}\frac{1}{j-1} \sim C\log m$$

as $m \to \infty$. $\square$

This allows us to prove the upper bound corresponding (qualitatively) to the lower bound in (4.17).

PROPOSITION 4.12. *For $c, K \in (0, \infty)$, the seed-bank coalescent satisfies*

$$(4.24) \qquad\qquad \limsup_{n \to \infty}\frac{\mathbb{E}[T_{\text{MRCA}}[n]]}{\log\log n} < \infty.$$

PROOF. Assume that the initial n individuals in the sample of the process $(\Pi_t^{(n)})_{t\geq 0}$ are labelled $1, \ldots, n$. Let

$$\mathcal{S}_r := \{k \in [n] : \exists 0 \leq t \leq r : k \text{ belongs to an } s\text{-block at time } t\}$$

denote those lines that visit the seed-bank at some time up to t. Let

$$\varrho^n := \inf\{r \geq 0 : |\mathcal{S}_r^c| = 1\}$$

be the first time that all those individuals which so far had not entered the seed-bank have coalesced. Note that ϱ^n is a stopping time for the process $(\Pi_t^{(n)})_{t\geq 0}$, and $N_{\varrho^n}^{(n,0)}$ and $M_{\varrho^n}^{(n,0)}$ are well-defined as the number of plant blocks, respectively, seed blocks of $\Pi_{\varrho^n}^{(n)}$. By a comparison of ϱ^n to the time to the most recent common ancestor of Kingman's coalescent cf. [28], $\mathbb{E}[\varrho^n] \leq 2$ for any $n \in \mathbb{N}$, and thus

$$
\begin{aligned}
\mathbb{E}[T_{\text{MRCA}}[(n,0)]] &\leq 2 + \mathbb{E}[T_{\text{MRCA}}[(N_{\varrho^n}^{(n,0)}, M_{\varrho^n}^{(n,0)})]] \\
&\leq 2 + \mathbb{E}[T_{\text{MRCA}}[(0, N_{\varrho^n}^{(n,0)} + M_{\varrho^n}^{(n,0)})]],
\end{aligned}
$$
(4.25)

where the last inequality follows from the fact that every seed has to become a plant before coalescing. Recall A^n from (4.16) and observe that

$$N_{\varrho^n}^{(n,0)} + M_{\varrho^n}^{(n,0)} \leq A^n + 1 \qquad \text{stochastically.}$$
(4.26)

This follows from the fact that for every individual, the rate at which it is involved in a coalescence is increased by the presence of other individuals, while the rate of migration is not affected. Thus, by coupling $(N_t, M_t)_{t\geq 0}$ to a system where individuals, once having jumped to the seed-bank, remain there forever, we see that $N_{\varrho^n} + M_{\varrho^n}$ is at most $A^n + 1$.

By the monotonicity of the coupling with $(\overline{N}_t, \overline{M}_t)$, we thus see from (4.25), for $\varepsilon > 0$,

$$
\begin{aligned}
\mathbb{E}[T_{\text{MRCA}}[n]] &\leq 2 + \mathbb{E}[\overline{T}_{\text{MRCA}}[A^n + 1]] \\
&= 2 + \mathbb{E}[\overline{T}_{\text{MRCA}}[A^n + 1]\mathbf{1}_{\{A^n \leq (2c+\varepsilon)\log n\}}] \\
&\quad + \mathbb{E}[\overline{T}_{\text{MRCA}}[A^n + 1]\mathbf{1}_{\{A^n > (2c+\varepsilon)\log n\}}].
\end{aligned}
$$
(4.27)

From Lemma 4.11, we obtain

$$\mathbb{E}[\overline{T}_{\text{MRCA}}[A^n + 1]\mathbf{1}_{\{A^n \leq (2c+\varepsilon)\log n\}}] \leq C\log(2c - \varepsilon)\log n \leq C\log\log n,$$

and since $A^n \leq n$ in any case, we get

$$\mathbb{E}[\overline{T}_{\text{MRCA}}[A^n + 1]\mathbf{1}_{\{A^n > (2c+\varepsilon)\log n\}}] \leq C\log n \cdot \mathbb{P}(A^n > (2c+\varepsilon)\log n) \leq C.$$

This completes the proof. $\square$

REMARK 4.13. In the same manner as in the proof of Theorem 4.6, one can show that for any $a, b \geq 0$,

$$\mathbb{E}[T_{\text{MRCA}}[an, bn]] \asymp \log(\log(an) + bn).$$

APPENDIX

PROPOSITION A.1. *Assume $c = \varepsilon N = \delta M$ and $M \to \infty$, $N \to \infty$. Let $(D_{N,M})_{N,M \in \mathbb{N}}$ be an array of positive real numbers. Then the discrete generator of the allele frequency process $(X^N_{\lfloor D_{N,M} t \rfloor}, Y^M_{\lfloor D_{N,M} t \rfloor})_{t \in \mathbb{R}^+}$ on time-scale $D_{N,M}$ is given by*

$$(A^N f)(x, y) = D_{N,M} \left[\frac{c}{N}(y - x)\frac{\partial f}{\partial x}(x, y) + \frac{c}{M}(x - y)\frac{\partial f}{\partial y}(x, y) \right.$$
$$\left. + \frac{1}{N}\frac{1}{2}x(1 - x)\frac{\partial^2 f}{\partial x^2}(x, y) + R(N, M) \right],$$

where the remainder term $R(N, M)$ satisfies that there exists a constant $C_1(c, f) \in (0, \infty)$, independent of N and M, such that

$$\left| R(N, M) \right| \le C_1 \left(N^{-3/2} + M^{-2} + N^{-1}M^{-1} + NM^{-3} \right).$$

In particular, in the situation where $M = O(N)$ as $N \to \infty$ and $D_{N,M} = N$ we immediately obtain Proposition 2.4.

PROOF OF PROPOSITION A.1. We calculate the generator of $(X^N_k, Y^M_k)_{k \ge 0}$ depending on the scaling $(D_{N,M})_{N,M \in \mathbb{N}}$. For $f \in \mathcal{C}^3([0, 1]^2)$ we use Taylor expansion in 2 dimensions to obtain

$$(A^N f)(x, y) = \frac{1}{D_{N,M}} \left[\frac{\partial f}{\partial x}(x, y)\mathbb{E}_{x,y}[X^N_1 - x] + \frac{\partial f}{\partial y}(x, y)\mathbb{E}_{x,y}[Y^M_1 - y] \right.$$

$$+ \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(x, y)\mathbb{E}_{x,y}[(X^N_1 - x)^2]$$

$$+ \frac{1}{2}\frac{\partial^2 f}{\partial y^2}(x, y)\mathbb{E}_{x,y}[(Y^M_1 - y)^2]$$

$$+ \frac{\partial^2 f}{\partial x\, \partial y}(x, y)\mathbb{E}_{x,y}[(X^N - x)(Y^M_1 - y)]$$

$$\left. + \mathbb{E}_{x,y}\left[\sum_{\substack{\alpha,\beta \in \mathbb{N}_0 \\ \alpha+\beta=3}} R^{\alpha,\beta}(X^N_1, Y^N_1)(X^M_1 - x)^\alpha (Y^M_1 - y)^\beta \right] \right],$$

where the remainder is given by

$$R^{\alpha,\beta}(\bar{x}, \bar{y}) := \frac{\alpha + \beta}{\alpha!\beta!} \int_0^1 (1 - t)^{\alpha+\beta-1} \frac{\partial^3 f}{\partial x^\alpha\, \partial y^\beta}(x - t(\bar{x} - x), y - t(\bar{y} - y))\, dt$$

for any $\bar{x}, \bar{y} \in [0, 1]$. In order to prove the convergence, we thus need to calculate or bound all the moments involved in this representation.

Given $\mathbb{P}_{x,y}$, the following holds: By Proposition 2.2,

$$X_1^N = \frac{1}{N}(U + Z),$$

$$Y_1^M = \frac{1}{M}(yM - Z + V),$$

in distribution where U, V and Z are independent random variables such that

$$U \sim \text{Bin}(N - c, x),$$

$$V \sim \text{Bin}(c, x),$$

$$Z \sim \text{Hyp}(M, c, yM).$$

Thus, we have

(A.1) $\qquad \mathbb{E}_{x,y}[U] = Nx - cx, \qquad \mathbb{E}_{x,y}[V] = cx, \qquad \mathbb{E}_{x,y}[Z] = cy,$

and moreover

$$\mathbb{V}_{x,y}(U) = (N - c)x(1 - x).$$

One more observation is that as $0 \leq V \leq c$ and $0 \leq Z \leq c$, it follows that $|Z - cX| \leq c$ and $|V - Z| \leq c$, which implies that for every $\alpha \in \mathbb{N}$

$$\left|\mathbb{E}_{x,y}\big[(Z - cX)^\alpha\big]\right| \leq c^\alpha,$$

$$\left|\mathbb{E}_{x,y}\big[(Z - V)^\alpha\big]\right| \leq c^\alpha,$$

and for every $\alpha, \beta \in \mathbb{N}$

(A.2) $\qquad\qquad \left|\mathbb{E}_{x,y}\big[(Z - cX)^\alpha (V - Z)^\beta\big]\right| \leq c^{\alpha+\beta}.$

We are now prepared to calculate all the mixed moments needed.

$$\begin{aligned}
\mathbb{E}_{x,y}[X_1^N - x] &= \frac{1}{N}\mathbb{E}_{x,y}[U + Z - Nx] \\
&= \frac{1}{N}\mathbb{E}_{x,y}[U - Nx + cx] + \frac{1}{N}\mathbb{E}_{x,y}[Z - cx] \\
&= \frac{c}{N}(y - x).
\end{aligned}$$

Here, we used (A.1), in particular $\mathbb{E}_{x,y}[U - Nx + cx] = \mathbb{E}_{x,y}[U - \mathbb{E}_{x,y}[U]] = 0$. Similarly,

$$\begin{aligned}
\mathbb{E}_{x,y}[Y_1^M - y] &= \frac{1}{M}\mathbb{E}_{x,y}[My + V - Z - My] \\
&= \frac{1}{M}\mathbb{E}_{x,y}[V - Z] \\
&= \frac{c}{M}(x - y).
\end{aligned}$$

Noting $X_1^N - x = \frac{1}{N}(U - Nx + cx) + \frac{1}{N}(Z - cx)$ leads to

$$
\begin{aligned}
\mathbb{E}_{x,y}[(X_1^N - x)^2] &= \frac{1}{N^2}\mathbb{E}_{x,y}[(U - Nx + cx)^2] \\
&\quad + \frac{2}{N^2}\mathbb{E}_{x,y}[U - Nx + cx]\mathbb{E}_{x,y}[Z - cx] \\
&\quad + \frac{1}{N^2}\mathbb{E}_{x,y}[(Z - cx)^2] \\
&= \frac{1}{N^2}\mathbb{V}_{x,y}[U] + \frac{1}{N^2}\mathbb{E}_{x,y}[(Z - cx)^2] \\
&= \frac{1}{N}x(1 - x) - \frac{c}{N^2}x(1 - x) + \frac{1}{N^2}\mathbb{E}_{x,y}[(Z - cx)^2],
\end{aligned}
$$

where

$$
\left| -\frac{c}{N^2}x(1 - x) + \frac{1}{N^2}\mathbb{E}_{x,y}[(Z - cx)^2] \right| \leq \frac{c^2}{N^2}.
$$

Moreover, we have

$$
\left| \mathbb{E}_{x,y}[(Y_1^M - y)^2] \right| = \left| \frac{1}{M^2}\mathbb{E}_{x,y}[(V - Z)^2] \right| \leq \frac{c^2}{M^2}.
$$

Using equation (A.2), we get

$$
\begin{aligned}
&\left| \mathbb{E}_{x,y}[(X_1^N - x)(Y_1^M - y)] \right| \\
&\quad \leq \left| \frac{1}{NM}\mathbb{E}_{x,y}[U - xN + cx]\mathbb{E}_{x,y}[V - Z] \right| \\
&\quad\quad + \left| \frac{1}{NM}\mathbb{E}_{x,y}[(Z - cx)(V - Z)] \right| \\
&\quad \leq \frac{c^2}{NM}.
\end{aligned}
$$

We are thus left with the task of bounding the remainder term in the Taylor expansion. Since $f \in C^3([0, 1]^2)$, we can define

$$
\tilde{C}^f := \max\left\{ \frac{\partial^3 f}{\partial x^\alpha \partial y^\beta}(\bar{x}, \bar{y}) \Big| \alpha, \beta \in \mathbb{N}_0, \alpha + \beta = 3, \bar{x}, \bar{y} \in [0, 1] \right\},
$$

which yields a uniform estimate for the remainder in the form of

$$
\left| R^{\alpha,\beta}(\bar{x}, \bar{y}) \right| \leq \frac{1}{\alpha!\beta!\tilde{C}^f},
$$

which in turn allows us to estimate

$$\left| \mathbb{E}_{x,y}\left[\sum_{\substack{\alpha,\beta\in\mathbb{N}_0 \\ \alpha+\beta=3}} R^{\alpha,\beta}(X_1^N, Y_1^N)(X_1^N - x)^\alpha (Y_1^M - y)^\beta \right] \right|$$

$$\leq \frac{1}{\alpha!\beta!\bar{C}^f} \sum_{\substack{\alpha,\beta\in\mathbb{N}_0 \\ \alpha+\beta=3}} \mathbb{E}_{x,y}\left[|(X_1^N - x)^\alpha (Y_1^M - y)^\beta| \right].$$

Thus the claim follows if we show that the third moments are all of small enough order in N and M. Observe that for $\alpha \in \{0, 1, 2\}$ we have

$$(A.3) \qquad \mathbb{E}_{x,y}\left[|(U - Nx + cx)|^\alpha \right] \leq N.$$

For $\alpha = 0$, this is trivially true, for $\alpha = 1$ it is due to the fact that the binomial random variable U is supported on $0, \dots, N - c$ and $Nx - cx$ is its expectation, and for $\alpha = 2$ it follows from the fact that $(U - Nx + cx)^2 = |(U - Nx + cx)^2|$ and the formula for the variance of a binomial random variable. For $\alpha = 3$, it follows, for example, from Lemma 3.1 in [14] that

$$(A.4) \qquad \mathbb{E}_{x,y}\left[|(U - Nx + cx)|^3 \right] = O(N^{3/2}).$$

Thus, we get for any $0 \leq \alpha, \beta \leq 3$ such that $\alpha + \beta = 3$ that

$$\mathbb{E}_{x,y}\left[|(X_1 - x)^\alpha (Y_1^M - y)^\beta| \right]$$

$$= \frac{1}{N^\alpha M^\beta} \sum_{i=0}^\alpha \binom{\alpha}{i} \mathbb{E}_{x,y}\left[|(U - Nx + cx)^i (Z - cx)^{\alpha-i} (V - Z)^\beta| \right]$$

$$\leq \frac{1}{N^\alpha M^\beta} \sum_{i=0}^\alpha \binom{\alpha}{i} \mathbb{E}_{x,y}\left[|(U - Nx + cx)^i| \right] \mathbb{E}_{x,y}\left[|(Z - cx)^{\alpha-i} (V - Z)^\beta| \right]$$

$$\leq \frac{1}{N^\alpha M^\beta} \sum_{i=0}^\alpha \binom{\alpha}{i} N(2c)^{\alpha-i+\beta} 1_{\{1,2,3\}}(\alpha) + \frac{3(2c)^3}{N^{3/2}} 1_{\{3\}}(\alpha)$$

$$\leq C\left(\frac{1}{NM} + \frac{1}{M^2} + \frac{1}{N^{3/2}} \frac{N}{M^3} \right),$$

from (A.2), (A.3) and (A.4), where the constant C depends only on c. This completes the proof. $\square$

REFERENCES

[1] BERESTYCKI, N. (2009). *Recent Progress in Coalescent Theory. Ensaios Matemáticos* **16**. Sociedade Brasileira de Matemática, Rio de Janeiro. MR2574323

[2] BLATH, J., GONZÁLEZ CASANOVA, A., ELDON, B. and KURT, N. (2014). Genealogy of a Wright Fisher model with strong seed-bank component. Preprint.

[3] BLATH, J., GONZÁLEZ CASANOVA, A., KURT, N. and SPANÒ, D. (2013). The ancestral process of long-range seed bank models. *J. Appl. Probab.* **50** 741–759. MR3102512

[4] DONG, R., GNEDIN, A. and PITMAN, J. (2007). Exchangeable partitions derived from Markovian coalescents. *Ann. Appl. Probab.* **17** 1172–1201. MR2344303

[5] ETHERIDGE, A. (2011). *Some Mathematical Models from Population Genetics. Lecture Notes in Math.* **2012**. Springer, Heidelberg. MR2759587

[6] ETHIER, S. N. and KURTZ, T. G. (2005). *Markov Processes*: *Characterization and Convergence*, 2nd ed. Wiley, New York.

[7] FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications*. *Vol. I*, 3rd ed. Wiley, New York. MR0228020

[8] FISHER, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford Univ. Press, London. MR1785121

[9] GOLDSCHMIDT, C. and MARTIN, J. B. (2005). Random recursive trees and the Bolthausen–Sznitman coalescent. *Electron. J. Probab.* **10** 718–745 (electronic). MR2164028

[10] GONZÁLEZ CASANOVA, A., AGUIRRE-VON WOBESER, E., ESPÍN, G., SERVÍN-GONZÁLEZ, L., KURT, N., SPANÒ, D., BLATH, J. and SOBERÓN-CHÁVEZ, G. (2014). Strong seed-bank effects in bacterial evolution. *J. Theoret. Biol.* **356** 62–70.

[11] HERBOTS, H. M. (1994). Stochastic models in population genetics: Genealogical and genetic differentiation in structured populations. Ph.D. dissertation, Univ. London.

[12] HERBOTS, H. M. (1997). The structured coalescent. In *Progress in Population Genetics and Human Evolution* (*Minneapolis*, *MN*, 1994). *IMA Vol. Math. Appl.* **87** 231–255. Springer, New York. MR1493030

[13] JANSEN, S. and KURT, N. (2014). On the notion(s) of duality for Markov processes. *Probab. Surv.* **11** 59–120. MR3201861

[14] JENKINS, P. A., FEARNHEAD, P. and SONG, Y. S. (2014). Tractable stochastic models of evolution for loosely linked loci. Available at arXiv:1405.6863.

[15] KAJ, I., KRONE, S. M. and LASCOUX, M. (2001). Coalescent theory for seed bank models. *J. Appl. Probab.* **38** 285–300. MR1834743

[16] KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* **13** 235–248. MR0671034

[17] LENNON, J. T. and JONES, S. E. (2011). Microbial seed banks: The ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology* **9** 119–130.

[18] LEVIN, D. A. (1990). The seed bank as a source of genetic novelty in plants. *Amer. Nat.* **135** 563–572.

[19] NEUHAUSER, C. and KRONE, S. M. (1997). The genealogy of samples in models with selection. *Genetics* **145** 519–534.

[20] NOTOHARA, M. (1990). The coalescent and the genealogical process in geographically structured population. *J. Math. Biol.* **29** 59–75. MR1083665

[21] NUNNEY, L. (2002). The effective size of annual plant populations: The interaction of a seed bank with fluctuating population size in maintaining genetic variation. *Amer. Nat.* **160** 195–204.

[22] PITMAN, J. (1999). Coalescents with multiple collisions. *Ann. Probab.* **27** 1870–1902. MR1742892

[23] SCHWEINSBERG, J. (2000). A necessary and sufficient condition for the Λ-coalescent to come down from infinity. *Electron. Commun. Probab.* **5** 1–11 (electronic). MR1736720

[24] TAKAHATA, N. (1988). The coalescent in two partially isolated diffusion populations. *Genet. Res.* **53** 213–222.

[25] TELLIER, A., LAURENT, S. J. Y., LAINER, H., PAVLIDIS, P. and STEPHAN, W. (2011). Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proc. Natl. Acad. Sci. USA* **108** 17052–17057.

[26] TEMPLETON, A. R. and LEVIN, D. A. (1979). Evolutionary consequences of seed pools. *Amer. Nat.* **114** 232–249.

[27] VITALIS, R., GLÉMIN, S. and OLIVIERE, I. (2004). When genes got to sleep: The population genetic consequences of seed dormancy and monocarpic perenniality. *Amer. Nat.* **163** 295–311.

[28] WAKELEY, J. (2009). *Coalescent Theory*. Roberts and Co, Greenwood Village, Colorado.

[29] WRIGHT, S. (1931). Evolution in Mendelian populations. *Genetics* **16** 97–159.

[30] ŽIVKOVIĆ, D. and TELLIER, A. (2012). Germ banks affect the inference of past demographic events. *Mol. Ecol.* **21** 5434–5446.

INSTITUT FÜR MATHEMATIK, SEKR. MA 7-5
TECHNISCHE UNIVERSITÄT BERLIN
STRASSE DES 17. JUNI 136
10623 BERLIN
GERMANY
E-MAIL: blath@math.tu-berlin.de
             adriangcs@hotmail.com
             kurt@math.tu-berlin.de
             wilkeber@math.tu-berlin.de