Seed bank Cannings graphs: How dormancy smoothes random genetic drift

Adrián González Casanova, Lizbeth Peñaloza and Arno Siri-Jégousse

Universidad Nacional Autónoma de México. Instituto de Matemáticas, Unidad Cuernavaca. Av. Universidad s/n Periferica, 62210 Cuernavaca, Morelos, México, and University of California at Berkeley. Department of Statistics. Berkeley, California, USA.
E-mail address: `adrian.gonzalez@im.unam.mx, gonzalez.casanova@berkeley.edu`

Universidad del Mar, campus Huatulco. Instituto de Investigación de Matemáticas y Actuaría. Carretera Federal No 200, Km 250, 70989 Santa María Huatulco, Oaxaca, México.
E-mail address: `lizbeth@huatulco.umar.mx`

Universidad Nacional Autónoma de México. Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas. Circuito Escolar 3000, C.U., 04510 Coyoacán, CDMX, México.
E-mail address: `arno@sigma.iimas.unam.mx`

Abstract. In this article, we introduce a random (directed) graph model for the simultaneous forwards and backwards description of a rather broad class of Cannings models with a seed bank mechanism. This provides a simple tool to establish a sampling duality in the finite population size, and obtain a path-wise embedding of the forward frequency process and the backward ancestral process. Further, it allows the derivation of limit theorems that generalize celebrated results by Möhle to models with seed banks, and where it can be seen how the effect of seed banks affects the genealogies. The explicit graphical construction is a new tool to understand the subtle interplay of seed banks, reproduction and genetic drift in population genetics.

1. Introduction

Cannings models and their modifications, along with their multiple merger genealogies are a major topic in mathematical population genetics Sagitov (1999); Möhle and Sagitov (2001); Schweinsberg (2003); Birkner et al. (2018); Freund (2020); González Casanova et al. (2022a); Siri-Jégousse and Wences (2022). Also, in the last decade, the study of dormancy (also called seed bank effect) received significant attention Kaj et al. (2001); Blath et al. (2013, 2016, 2020). One of the unifying themes in both modeling areas is that they arise from extensions of the Wright-Fisher model, and that classical evolutionary forces such as genetic drift and selection are affected in important ways. While the theory of Cannings models is now robust, the study of models with dormancy is still work

in progress. The main goal of this paper is to stabilize a framework in which seed banks can be combined with Cannings models, and to generalize the known limiting results for models without dormancy to Cannings models with seed bank.

An important tool in population genetics is the moment duality for Markov processes. This technique establishes a mathematical relation between forward and backward in time processes. The celebrated duality between the Wright-Fisher diffusion and the Kingman coalescent was gradually generalized to a wide class of neutral population genetics models, including some finite size discrete populations such as Cannings-type models González Casanova and Spanò (2018). In this situation, the duality leads to asymptotic results for both forward frequency and genealogical processes. In the context of dormancy, duality was established for the seed bank diffusion, which arises as a limit in models with geometric seed bank Blath et al. (2016). However, for discrete seed bank models, the duality relation is the first open gap that this article aims to close.

There are two main models for dormancy phenomena.

- Kaj et al. The model defined in Kaj et al. (2001) is based on the Wright-Fisher model with additional multi-generational jumps of (bounded) size, the system has been extended to geometric jump sizes of bounded expected range in Koopmann et al. (2017) (which also provide some insight into the forward in time frequency diffusion), to the general finite expectation case in Blath et al. (2013), and even to unbounded (heavy-tailed) jump sizes in Blath et al. (2015).
- Blath et al. A second modeling frame is given by an external seed bank in terms of a "second island" (in the spirit of Wright's island model), effectively leading to geometric jump sizes on the evolutionary scale. Here, forward and backward limits have been constructed, giving rise to the seed bank diffusion and the seed bank coalescent Blath et al. (2016) (see more analysis and generalization in González Casanova et al. (2022b); Blath et al. (2020) and an interesting connection with metapopulations in Lambert and Ma, 2015).

Both modeling frames (generational jumps and second island) have their advantages and disadvantages. For the Wright-Fisher model with multi-generational jumps, one typically loses the Markov property. For the island version, one retains the Markov property but then needs to investigate two-dimensional frequency processes, which in the limit are harder to analyze than one-dimensional diffusions, since e.g. the Feller theory is missing (this can in part be replaced by recent theory for polynomial diffusions Blath et al., 2019). Interestingly, it turns out that for the limiting frequency processes, both approaches can be two sides of the same medal.

In none of the above approaches, more general reproductive mechanisms, such as based on Cannings models, have been analyzed. This paper's second aim is to close this gap. We present an extended framework for the simultaneous construction of seed bank models with general multi-generational jump distributions and Cannings-type reproductive laws satisfying a paintbox construction. We are also able to obtain forward and backward convergence results (extending Kaj et al. (2001), Koopmann et al. (2017) and Blath et al., 2013) and to provide an explicit sampling duality, which is valid already in the finite individual models.

More precisely, we show that if a sequence of Cannings models (with no seed bank effect) is in the universality class of the Kingman coalescent, meaning that its ancestral process converges in the evolutionary scale to the Kingman coalescent, then the ancestry of the same sequence with a seed bank effect will converge to the Kingman coalescent delayed by a constant β^2, where $\beta < \infty$ is the expected number of generations that separates an individual from its ancestor. This extends the results of Kaj et al. (2001) and Blath et al. (2013). Convergence of the frequency process to the solution of the Wight-Fisher diffusion with the same delay is also proved. We go further and study how sequences of seed bank models with divergent expectations can make sequences of Cannings models that originally were not in the Kingman class, converge to the Kingman coalescent. This is achieved using the mixing time of some auxiliary Markov chains introduced in Kaj et al. (2001).

If instead of considering Cannings processes in the Kingman class we consider that their genealogy converges to a Ξ-coalescent, we show that their seed bank modification converges to a Ξ^β-coalescent. Heuristically, the transformation $\Xi \to \Xi^\beta$ consists in dividing by β all the non-dust boxes in a Ξ paintbox event to obtain a Ξ^β paintbox event. Similar asymptotics are shown for the forward process. All those results are extended for models in the presence of mutations.

Note that the interplay of general reproduction and seed banks with other evolutionary forces can be subtle, and we provide a framework for its analysis (also regarding the real-time embedding of coalescent-based estimates, see e.g. Blath et al., 2020).

The paper is organized as follows. In section 2 we construct a random graph that allows us to embed the ancestry and the frequency processes of both Cannings and dormancy models simultaneously and study the duality relation of the processes forward and backward in time. Furthermore, we analyze the scaling limits of the ancestral process in presence of skewed reproduction mechanisms and dormancy. We give conditions for convergence to the Kingman coalescent and study scenarios beyond this universality class, where we can describe how seed bank phenomena reduce the typical size coalescence events when combining seed banks with Cannings models that would, in absence of the seed bank component, converge to a Λ- or a Ξ- coalescent. Section 3 uses the moment duality to formally prove convergence of the frequency process to a Wright-Fisher diffusion. This intuitively clear result was missing in the literature, probably since the lack of Markov property for the frequency process makes usual techniques fail. In section 4 we study a variant of the seed bank random graph where mutations are added and we extend the results obtained in sections 2 and 3.

2. A random graph version of the model of Kaj, Krone and Lascoux

Consider a discrete-time haploid population of constant size $N \geq 1$ at each generation. The vertex set $V^N = \mathbb{Z} \times [N]$ represents the whole population. For each individual $v \in V^N$, denote by $g(v)$ its generation and by $\ell(v)$ its label so that $v = (g(v), \ell(v))$. We denote the g-th generation of the population by $V_g^N := \{v \in V^N : g(v) = g\}$. Set a probability measure $\mathcal{W}^N$ on the exchangeable probability measures on $[N]$. Let $\{\bar{W}_g^N\}_{g\in\mathbb{Z}}$ be a sequence of independent $\mathcal{W}^N$-distributed random variables with $\bar{W}_g^N = \{W_v^N\}_{v\in V_g^N}$. Each variable W_v^N gives the reproductive weight of the individual v in the population graph. This multinomial setting can be extended to some more general Cannings models (as in Möhle and Sagitov, 2001) or non-exchangeable reproductive success (as in Siri-Jégousse and Wences, 2022). Also, consider a sequence $\{m_N\}_{N\geq 1}$ of integers and set a probability measure μ^N on $[m_N]$. Let $\{J_v^N\}_{v\in V^N}$ be a collection of independent μ^N-distributed random variables. The variable J_v^N says how many generations ago an individual v's mother is living. Finally, set a collection of random variables in $[N]$, $\{U_v^N\}_{v\in V^N}$ such that U_v^N is the label of the mother of v. Its conditional distribution is

$$\mathbb{P}(U_v^N = k | J_v^N = j, \{\bar{W}_g^N\}_{g\in\mathbb{Z}}) = W_{(g(v)-j,k)}^N.$$

Definition 2.1. (The seed bank random di-graph) Consider the random set of directed edges

$$E^N = \{(v, (g(v) - J_v^N, U_v^N)), \text{ for all } v \in V^N\}.$$

The *seed bank random di-graph* with parameters N, $\mathcal{W}^N$ and μ^N is given by $G^N := (V^N, E^N)$.

Two classical examples are

- the Kaj, Krone and Lascoux (KKL) seed bank graph Kaj et al. (2001), in this case μ^N has finite support $[m]$, i.e. $m_N = m$, and $\mathcal{W}^N = \delta_{(1/N,\ldots,1/N)}$.
- the Cannings model with parameter $\mathcal{W}^N$ Cannings (1974, 1975); Möhle and Sagitov (2001), in this case $\mu^N = \delta_1$.
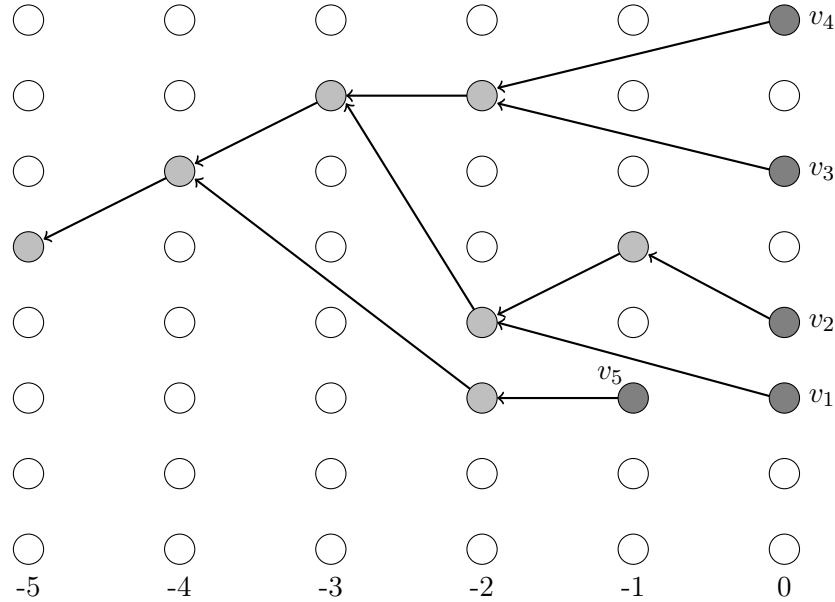
FIGURE 2.1. In this case $N = 8$ and $m_N = 2$. The gray circles represent the members of $S_0 = \{v_1, v_2, v_3, v_4, v_5\}$ where, for example, $v_2 = (0, 4)$ and $v_5 = (-1, 3)$. The light gray circles represent the ancestors of the sample. $\bar{A}_0^8 = (4, 1)$, $\bar{A}_1^8 = (2, 2)$, $\bar{A}_2^8 = (3, 0)$, $\bar{A}_3^8 = (1, 1)$, $\bar{A}_4^8 = (1, 0)$, $\bar{A}_5^8 = (1, 0)$.

For every $u, v \in V^N$ we denote by $\delta(u, v)$ the distance of u and v in the graph G^N, i.e. the number of vertices in a path from u to v or from v to u. Now let us define the ancestral process associated with this graph.

Definition 2.2 (The ancestral process). Fix a generation g_0 and S_{g_0} consisting in a sample of individuals living between generation g_0 and $g_0 - m_N + 1$, i.e. $S_{g_0} \subset \cup_{i=1}^{m_N} V_{g_0+1-i}^N$. For every $g \geq 0$, let $\mathcal{A}_g^N$ be the set composed by the most recent ancestors of the individuals of S_{g_0} that live at a generation $g_0 - g'$ for some $g' \geq g$, that is

$$\mathcal{A}_g^N = \{v \in \cup_{g'=g}^\infty V_{g_0-g'}^N : \exists u \in S_{g_0} \text{ such that } \delta(u, v) \leq \delta(u, v') \text{ for all } v' \in \cup_{g'=g}^\infty V_{g_0-g'}^N\}.$$

Define, for all $i \in [m_N]$,

$$A_g^{N,i} = |\mathcal{A}_g^N \cap V_{g_0-g+1-i}^N|$$

and $\bar{A}_g^N = (A_g^{N,1}, \ldots, A_g^{N,m_N})$. We call $\{\bar{A}_g^N\}_{g \geq 0}$ the ancestral process. In the sequel, we consider the initial configuration $S_{g_0}(\bar{n})$, for $\bar{n} = (n_1, \ldots, n_{m_N})$, such that $n_i \geq 0$ individuals are uniformly sampled (with repetition) from generation $g_0 + 1 - i$. We denote the law of the ancestral process of this sample by $\mathbb{P}_{\bar{n}}$. See Figure 2.1 for an illustration.

For simplicity, we suppose that $\sup\{i \geq 1 : n_i > 0\}$ does not depend on N. This model was introduced, for reproductions as in the Wright-Fisher model, by Kaj et al. (2001) directly, in the sense that they construct a random graph only implicitly. Our construction permits to provide a transparent relation between the ancestral process and the forward frequency process defined in section 3. Observe that $\{\bar{A}_g^N\}_{g \geq 0}$ is a Markov chain. We start our results by formalizing the remark on p. 290 in Kaj et al. (2001). This illustrative result is established when the Cannings model is in the domain of attraction of the Kingman coalescent, although it can be easily generalized to any type of reproduction law. Here we use the classical notations c_N (resp. d_N) that denote the probability that two (resp. three) given individuals choose the same parent in a Cannings model.

Those notations will be helpful all along the paper. Recall, e.g. from Möhle and Sagitov (2001), that the genealogies of a Cannings model fall into the domain of attraction of the Kingman coalescent when $c_N \to 0$ and $d_N = o(c_N)$ while multiple merger coalescents arise when d_N and c_N are of the same order.

Proposition 2.3 (Reformulation of Theorem 1 in Kaj et al., 2001). *Suppose that*

$$c_N = N\mathbb{E}[(W_v^N)^2] \to 0 \text{ and } d_N = N\mathbb{E}[(W_v^N)^3] = o(c_N).$$

Let $M(n)$ be a multinomial random variable with parameters n and $\{\mu^N(i)\}_{i=1}^{m_N}$. Also, for any $\bar{n} = (n_1, \ldots, n_{m_N}) \in [N]^{m_N}$, let $Z(\bar{n}) = (n_2, \ldots, n_{m_N}, 0) + M(n_1)$. Then, the transitions of $\{\bar{A}_g^N\}_{g \geq 0}$ can be written in terms of M and Z as follows.

- $\mathbb{P}_{\bar{n}}(\bar{A}_1^N = Z(\bar{n})) = 1 - \sum_{i=1}^{\infty} c_N[\binom{n_1}{2}\mu^N(i)^2 + \mu^N(i)n_1 n_{i+1}] + o(c_N)$
- $\mathbb{P}_{\bar{n}}(\bar{A}_1^N = Z(\bar{n}) - e_i) = c_N[\binom{n_1}{2}\mu^N(i)^2 + \mu^N(i)n_1 n_{i+1}] + o(c_N)$

where e_i is the vector with the i-th coordinate equal to 1 and the others equal to 0, for all $i \geq 1$.

Proof: We need to make two observations. First note that all the randomness in the transitions of the chain $\{\bar{A}_g^N\}_{g \geq 0}$ lies in what happens to the first coordinate. If for some $g \geq 0$, $\bar{A}_g^N = (0, n_2, \ldots, n_{m_N})$ it is easy to see that $\bar{A}_{g+1}^N = (n_2, \ldots, n_{m_N}, 0)$ almost surely. On the other hand, if $n_1 > 0$, the individuals that are in $\mathcal{A}_g^N \cap V_{g_0-g}^N$ cannot belong to $\mathcal{A}_{g+1}^N$. Then, each of these individuals, if denoted by v, must be replaced by an individual which lives J_v^N generations in the past, that is

$$\mathbb{P}_{e_1}(\bar{A}_1^N = e_i) = \mu^N(i).$$

Further, if $n_1 > 1$, one needs to find n_1 new ancestors, but some of them could be the same due to some coalescence. The complete picture is as follows. For $i \geq 2$ and $j, k \geq 0$, and by denoting e_0 for the null vector,

$$\mathbb{P}_{2e_1+e_i}(\bar{A}_1^N = e_{i-1} + e_j + e_k) = \begin{cases} 2\mu^N(j)\mu^N(k) & \text{if } i-1 \neq j \neq k \\ (\mu^N(j))^2(1-c_N) & \text{if } i-1 \neq j, j = k \\ 2\mu^N(i-1)\mu^N(k)(1-c_N) & \text{if } i-1 = j, j \neq k \\ (2\mu^N(i-1)\mu^N(j) + (\mu^N(j))^2)c_N & \text{if } i-1 \neq j, k = 0 \\ (\mu^N(i-1))^2(c_N - d_N) & \text{if } i-1 = j, k = 0 \\ (\mu^N(i-1))^2 d_N & \text{if } j = k = 0 \end{cases} \quad . \quad (2.1)$$

The proof follows easily after these observations. □

We now construct a less natural backward process which will be very useful when establishing its moment duality with the forward process in section 3. We start by defining it in a graphical and intuitive way. More formal definitions will follow all along the section.

Definition 2.4 (The window process). Fix a generation g_0, and $S_{g_0} \subset \cup_{i=1}^{m_N} V_{g_0+1-i}^N$. In the genealogical tree of the sample S_{g_0}, define the variable $B_g^{N,1}$ as the number of edges arriving to generation $g_0 - g$ (plus the number of individuals of S_{g_0} living at this generation). For any $i \in \{2, m_N\}$, let $B_g^{N,i}$ be the number of edges crossing generation $g_0 - g$ and arriving to generation $g_0 - g - i + 1$ (plus the number of individuals of S_{g_0} living at this generation). Then, let $\bar{B}_g^N := (B_g^{N,1}, \ldots, B_g^{N,m_N})$. We call $\{\bar{B}_g^N\}_{g \geq 0}$ the window process. As for the ancestral process, we denote by $\mathbb{P}_{\bar{n}}$ the law of the window process generated from the initial sample $S_{g_0}(\bar{n})$.

In Figure 2.1, the values of the window process are $\bar{B}_0^8 = (4, 1)$, $\bar{B}_1^8 = (2, 3)$, $\bar{B}_2^8 = (5, 0)$, $\bar{B}_3^8 = (2, 1)$, $\bar{B}_4^8 = (2, 0)$, $\bar{B}_5^8 = (1, 0)$. The window process and the ancestral process only differ in the time where we acknowledge a coalescence event. In the window process coalescence events only

occur in the first coordinate, while in the ancestral process coalescence events may take place at every entry (see Figure 2.1).

The following equivalent (in law) definition of the window process allows us to compare it with the ancestral process. Let $C^N(n)$ be the number of ancestors after one generation of a sample of n individuals in a Cannings model with weights distributed as $\mathcal{W}^N$. As in Proposition 2.3, let $M(n)$ be a multinomial random variable with parameters n and $\{\mu^N(i)\}_{i=1}^{m_N}$. Given $\bar{B}_{g-1}^N = \bar{n} = (n_1, \ldots, n_{m_N}) \in [N]^{m_N}$,

$$\bar{B}_g^N = (n_2, \ldots, n_{m_N}, 0) + M(C^N(n_1)).$$

in distribution. It is left to the reader to show that indeed both definitions are equivalent.

The process $\{\bar{B}_g^N\}_{g \geq 0}$ can be expressed in terms of a particle system. Fix N, μ^N and $\mathcal{W}^N$. Let $Y_g^N = (R_g^N, L_g^N)$ define a Markov chain with state space $\mathbb{Z}_+ \times [N]$ and transition probabilities, conditional on the weights $\bar{W}_g^N$,

$$\mathbb{P}\big(Y_g^N = (i,k)|\{\bar{W}_g^N\}_g, Y_{g-1}^N = (1,j)\big) = W_{(g_0-g+1-i,k)}^N \mu^N(i)$$

for every $i \geq 1$, $k \in [N]$, and

$$\mathbb{P}\big(Y_g^N = (i,k)|\{\bar{W}_g^N\}_g, Y_{g-1}^N = (i+1,j)\big) = W_{(g_0-g,k)}^N$$

for every $i \geq 1$ and $k \in [N]$.

Proposition 2.5. *Set $n = \sum n_i$ to be the total size of the initial sample. For every $g \geq 0$, consider n (conditional on $\{\bar{W}_g^N\}_g$) independent realizations of Y_g^N, that we call $Y_g^{N,j} = (R_g^{N,j}, L_g^{N,j})$ for $1 \leq j \leq n$. Let $\sigma^{N,1} = \infty$ and*

$$\sigma^{N,j} = \inf\{g \geq 1 : Y_g^{N,j} = Y_g^{N,j'} = (1,k), \text{ for some } j' < j \text{ such that } \sigma^{N,j'} > g, k \in [N]\}.$$

For all $i \geq 1$, set $\sum_{j=1}^n 1_{\{R_0^{N,j}=i\}} = B_0^{N,i}$. Then, the i-th component $B_g^{N,i}$ of the random vector $\bar{B}_g^N$ is equal in distribution to $\sum_{j=1}^n 1_{\{R_g^{N,j}=i\}} 1_{\{\sigma^{N,j} \geq g\}}$, for all $g \geq 0$.

Proof: The proof consists in observing that $g_0 - R_g^{N,j} - g + 1$ is equal in distribution to the generation of the most recent ancestor, living at a generation $g_0 - g'$ for some $g' \geq g$, of a fixed individual in the initial sample S_{g_0}. So we couple these two processes. At the particular times in which $R_g^{N,j} = 1$ (and thus a coalescence event can occur in the window process) we take $L_g^{N,j}$ to be the label of the closest ancestor. Then $\sigma^{N,j}$ corresponds to the generation at which individual j's ancestral lineage is involved into a coalescence event with the ancestral lineage of an individual of lower level. Under this coupling,

$$B_g^{N,i} = \sum_{j=1}^n 1_{\{R_g^{N,j}=i\}} 1_{\{\sigma^{N,j} \geq g\}}$$

almost surely. $\qquad \square$

At this point it could seem unnecessary that the process L jumps at every time since it looks like it only has a role when R is reaching 1. However this independent construction of R and L will be important in the proof of Theorem 2.6. The chain $\{Y_g^N\}_{g \geq 0}$ provides a very convenient coupling to the ancestral and the window processes, mainly because $\{R_g^N\}_{g \geq 0}$ has an invariant measure given by

$$\nu^N(i) = \frac{\mathbb{P}(J_v^N \geq i)}{\mathbb{E}[J_v^N]}. \tag{2.2}$$

To see this, just observe that the chain has two types of behaviors. Using the notation $\mathbb{P}_j(\cdot) = \mathbb{P}(\cdot|R_0^N = j)$, we have

(1) Deterministic transitions: if $j > 1$, then $\mathbb{P}_j(R_1^N = j - 1) = 1$

(2) Random transitions: for $j \geq 1$, $\mathbb{P}_1(R_1^N = j) = \mathbb{P}(J_v^N = j) = \mu^N(j)$.

Then,

$$\sum_{j=1}^{\infty} \mathbb{P}_j(R_1^N = i)\nu^N(j) = \mathbb{P}_{i+1}(R_1^N = i)\nu^N(i+1) + \mathbb{P}_1(R_1^N = i)\nu^N(1)$$

$$= \frac{\mathbb{P}(J_v^N \geq i+1)}{\mathbb{E}[J_v^N]} + \frac{\mathbb{P}(J_v^N = i)}{\mathbb{E}[J_v^N]}$$

$$= \nu^N(i).$$

Our way to compare two Markov chains consists in applying coupling concepts developed in Levin and Peres (2017). Let us first recall the definition of mixing time (see page 55 of Levin and Peres, 2017). We denote $d^N(g) = \max_{j \in [m_N]} ||\mathbb{P}_j(R_g^N \in \cdot) - \nu^N(\cdot)||_{TV}$ and the mixing time $\tau_N = \inf\{g \geq 0 : d^N(g) < 1/4\}$. We allow $\tau_N = 0$, because this is true in the important case $\mathbb{P}(J_v^N = 1) = 1$, which is the case without seed bank.

The main theorem of Kaj et al. (2001) (proved for μ^N with finite support and extended to finite expectation in Blath et al., 2013) consists in showing that the L_1 norm of the ancestral process converges weakly to the block counting process of the Kingman coalescent under a constant time change. Here we extend this result to the window process and to some more general Cannings' mechanism.

Theorem 2.6 (Convergence of the window process I: Kingman limit). *Consider a seed bank digraph with parameters $N, \mathcal{W}^N$ and μ^N. Suppose that $\beta_N = \mathbb{E}[J_v^N] < \infty$. Let τ_N be the mixing time of $\{R_g^N\}_{g \geq 0}$, $c_N = N\mathbb{E}[(W_v^N)^2]$ and $d_N = N\mathbb{E}[(W_v^N)^3]$. Assume that $\mu^N(1) > 0$ and that*

$$c_N/\beta_N^2 \to 0, \quad N^\varepsilon \tau_N c_N \to 0, \quad (1/4)^{N^\varepsilon}\beta_N^2 \to 0 \quad and \quad d_N/(\beta_N c_N) \to 0. \tag{2.3}$$

for some $\varepsilon > 0$. Then, let $\{\bar{B}^N\}_{N \geq 1}$ be the sequence of window processes with parameters N and μ^N and starting condition $\bar{B}_0^N = \bar{n}$ for all $N \in \mathbb{Z}_+$ big enough. Then,

$$\lim_{N \to \infty} \{|\bar{B}_{\lfloor t\beta_N^2/c_N \rfloor}^N|\}_{t \geq 0} = \{N_t^K\}_{t \geq 0} \tag{2.4}$$

in the finite dimension sense, where $\{N_t^K\}_{t \geq 0}$ stands for the block counting process of a Kingman coalescent.

Furthermore, suppose that ν^N converges to a measure ν as $N \to \infty$. Let $V^{t,K}$ be a (conditional) multinomial random variable with parameters N_t^K and ν. For any fixed time $t > 0$, in distribution,

$$\lim_{N \to \infty} \bar{B}_{\lfloor t\beta_N^2/c_N \rfloor}^N = V^{t,K}. \tag{2.5}$$

Note that when $\beta_N \to \beta < \infty$ the third condition of Theorem 2.6 is automatically fulfilled. On the other side, when $\beta_N \to \infty$, then the fourth condition is always fulfilled because $d_N/c_N \leq 1$. The latter reflects the fact that a stronger seed bank effect (in the sense that the expected number of generations separating an individual from its ancestor tends to infinity) makes impossible the existence of multiple mergers. Theorem 2.8 below discusses the interplay between seed banks and random genetic drift. Note also that the second condition implies that $\tau_N < \infty$ for all but finitely many N, meaning that the support of μ^N is finite for all but at most finitely many N. In general, a process with finite mixing time can have an infinite support. In our case, a particle that starts in k takes k deterministic steps before jumping to a random location. This enforces that the support is a lower bound for the mixing time.

Remark 2.7. The arguments that we use in the proof of the theorem establish a convergence result in the finite dimension sense for (2.4). However this result can be strengthened into a convergence in distribution thanks to classical limit theorems, see for example Theorem 17.25 in Kallenberg (2021). This will also be the case in the forthcoming Theorems 2.8, 3.4, 4.7, 4.9 and 4.10.

Recall from the seminal work of Schweinsberg (2001) that Ξ-coalescents are characterized by a measure on Δ, the infinite simplex on $[0, 1]$. Every realization of this measure Ξ describes the coalescence rule at every jump time of the process. Let $F_\beta : \Delta \mapsto \Delta$ be such that for $A \subset \Delta$, $F_\beta(A) = \{\bar{y}/\beta, \bar{y} \in A\}$. To any finite measure Ξ over Δ, we associate a finite measure Ξ^β defined by the rule

$$\Xi^\beta(A) := \Xi(F_{1/\beta}(A) \cap \Delta)$$

for any borelian $A \subseteq \frac{1}{\beta}\Delta$, or equivalently

$$\Xi^\beta(F_\beta(B)) = \Xi(B)$$

for any borelian B on Δ. Observe that Ξ^β associates no weight on mass partitions $\bar{y} = (y_1, y_2, \dots)$ such that $\sum y_i > 1/\beta$. It is also remarkable that the seed bank effect sends the class of Λ-coalescents into itself. As an example, the $Beta(2 - \alpha, \alpha)$ coalescent with the characteristic measure, restricted to $(0, 1)$, $\Xi(A) = \int_A \frac{1}{\Gamma(2-\alpha)\Gamma(\alpha)} x^{1-\alpha}(1 - x)^{\alpha-1}dx$ turns to a coalescent with measure, now restricted to $(0, 1/\beta)$, $\Xi^\beta(A) = \int_A \frac{1}{\Gamma(2-\alpha)\Gamma(\alpha)} x^{1-\alpha}(\frac{1}{\beta} - x)^{\alpha-1}dx$.

Theorem 2.8 (Convergence of the window process II: Ξ limit). *Consider a seed bank di-graph with parameters $N, \mathcal{W}^N$ and μ^N. Suppose that $\beta_N = \mathbb{E}[J_v^N] \to \beta < \infty$. Assume that the ancestral process of a Cannings model driven by $\mathcal{W}^N$, that we denote by $\{C_g^N\}_{g\geq 0}$ is such that*

$$\lim_{N\to\infty} \{C_{\lfloor t/c_N \rfloor}^N\}_{t\geq 0}) = \{N_t^\Xi\}_{t\geq 0} \tag{2.6}$$

where $\{N_t^\Xi\}_{t\geq 0}$ stands for the block counting process of a Ξ-coalescent. Then

$$\lim_{N\to\infty} \{|\bar{B}_{\lfloor t/c_N \rfloor}^N|\}_{t\geq 0} = \{N_t^{\Xi^\beta}\}_{t\geq 0} \tag{2.7}$$

in the finite dimension sense.

Furthermore, suppose that ν^N converges to a measure ν as $N \to \infty$. Let V^{t,Ξ^β} be a (conditional) multinomial random variable with parameters $N_t^{\Xi^\beta}$ and ν. For any fixed time $t > 0$, in distribution,

$$\lim_{N\to\infty} B_{\lfloor t/c_N \rfloor}^N = V^{t,\Xi^\beta}. \tag{2.8}$$

Note that the rates of the Ξ^β-coalescent converge to those of the Kingman coalescent when $\beta \to \infty$, converting the Kingman coalescent to a limit model when β becomes large. This intuitively coincides with the hypothesis on β in Theorem 2.6.

Remark 2.9. The site frequency spectrum (SFS) provides a convenient way to visualize the transformations induced by the mapping Ξ^β of Theorem 2.8 to the shape of the coalescent. The SFS is a vector proportional to the branch lengths $(L_1, \dots, L_{n-1})$ of the coalescent tree, where L_i stands for the length of the lineages with exactly i leaves in the original sample of size n. In Figure 2.2 we see how the branch lengths are increased in a non linear way in the case of the Bolthausen-Sznitman coalescent. Note that when the genealogies are in the domain of attraction of the Kingman coalescent (Theorem 2.6), every coordinate of the SFS will be multiplied by the constant β.

Proof of Theorem 2.6: The proof consists in coupling the window process $\{\bar{B}_g^N\}_{g\geq 0}$ to a process which is "always in stationarity". If we suppose that $\{\bar{B}_g^N\}_{g\geq 0}$ starts a.s. with one lineage, i.e. $\bar{B}_0^N = e_k$ for some k, it is straightforward that it has a stationary distribution $\bar{\nu}^N$ given by $\bar{\nu}^N(e_i) = \nu^N(i)$. Now, let $\underline{Y}_g^{N,j} = (\underline{R}_g^{N,j}, L_g^{N,j})$ where $\{\underline{R}_g^{N,j}\}_{g\geq 0}$ is a sequence of independent ν^N-distributed random variables. Let

$$\underline{\sigma}^{N,j} = \inf\{g \geq 1 : \underline{Y}_g^{N,j} = \underline{Y}_g^{N,j'} = (1, k), \text{ for some } j' < j \text{ such that } \underline{\sigma}^{N,j'} > g, k \in [N]\}.$$
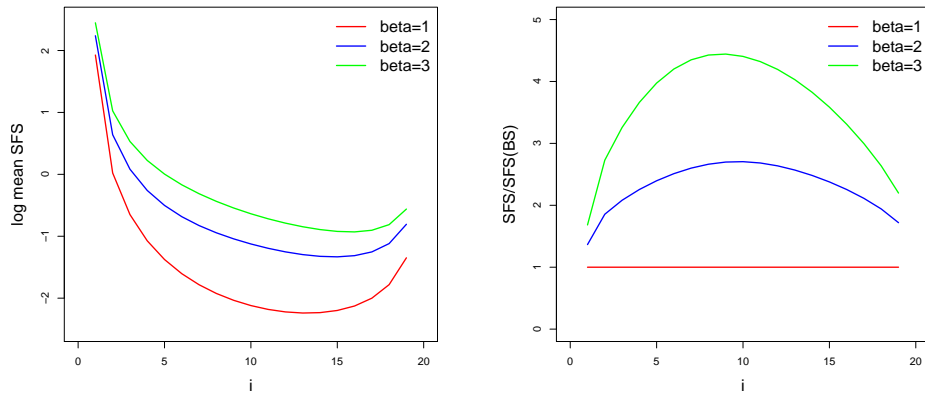
FIGURE 2.2. Expected SFS of the Ξ^β-coalescent where Ξ is the uniform measure on $(0, 1)$ (Bolthausen-Sznitman coalescent) and $\beta = 1, 2, 3$. The initial sample size is $n = 20$. The values are obtained thanks to computational methods based on phase-type theory Hobolth et al. (2019). An exact formula for $\beta = 1$ can be found in Kersting et al. (2021). Left: Expected SFS in logarithmic scale. Right: Expected SFS divided by the expected SFS of the Bolthausen-Sznitman coalescent.

Hence, inspired by the coupling of Proposition 2.5, we define an artificial window process by

$$\bar{Z}_g^N = (Z_g^{N,1}, \ldots, Z_g^{N,m_N})$$

where, as $n = \sum n_i$,

$$Z_g^{N,i} = \sum_{j=1}^{n} 1_{\{\underline{R}_g^{N,j} = i\}} 1_{\{\underline{\sigma}^{N,j} \geq g\}}.$$

The process $\{|\bar{Z}_g^N|\}_{g \geq 0}$ is Markovian.

We now proceed in two steps to prove (2.4). First, we calculate the generator of $\{|\bar{Z}_g^N|\}_{g \geq 0}$ in order to discover its scaling limit. Let $f : \mathbb{Z}_+ \to \mathbb{R}$ be a bounded function. Then

$$\mathcal{G}^N f(n) = \mathbb{E}[f(|\bar{Z}_1^N|) - f(n)] = \mathbb{P}(|\bar{Z}_1^N| = n - 1)[f(n-1) - f(n)] + O(\mathbb{P}(|\bar{Z}_1^N| = n - 2))$$

$$= \binom{n}{2} c_N (\nu^N(1))^2 [f(n-1) - f(n)] + O(d_N(\nu^N(1))^3)$$

$$= \frac{c_N}{\beta_N^2} \left[\binom{n}{2} [f(n-1) - f(n)] + O\left(\frac{d_N}{\beta_N c_N}\right) \right].$$

So we conclude that

$$\lim_{N \to \infty} \{|\bar{Z}_{\lfloor \beta_N^2 t/c_N \rfloor}^N|\}_{t \geq 0} = \{N_t^K\}_{t \geq 0} \tag{2.9}$$

in the finite dimension sense.

Second, let us couple $\{|\bar{Z}_{\lfloor \beta_N^2 t/c_N \rfloor}^N|\}_{t \geq 0}$ and $\{|\bar{B}_{\lfloor \beta_N^2 t/c_N \rfloor}^N|\}_{t \geq 0}$ to show that the same limit holds for the rescaled window process. The coupling consists in constructing for every $i \geq 1$ the random variable $(\underline{R}_{\rho_i}^{N,1}, \ldots, \underline{R}_{\rho_i}^{N,m_N})$ as the optimal coupling (defined in Remark 4.8, of Levin and Peres, 2017) of $(R_{\rho_i}^{N,1}, \ldots, R_{\rho_i}^{N,m_N})$, which depends on the initial condition $\bar{n} \in [N]^{m_N}$, and the stationary distribution $(\nu^N)^{\otimes m_N}$, where the times $\{\rho_i\}_{i \geq 1}$ correspond to the times where the processes $\{|\bar{Z}_{\lfloor \beta_N^2 t/c_N \rfloor}^N|\}_{t \geq 0}$ and $\{|\bar{B}_{\lfloor \beta_N^2 t/c_N \rfloor}^N|\}_{t \geq 0}$ can jump. More precisely, if we denote for any $p, q \in [n]$, $\rho_k^{N,p,q} = \inf\{g > \rho_{k-1}^{N,p,q} : L_g^{N,p} = L_g^{N,q}\}$ (with $\rho_0^{N,p,q} = 0$), then $\rho_i = \inf\{g > \rho_{i-1} : g = \rho_k^{N,p,q}$ for

some $p, q \in [n]$ with $p \neq q$, and some $k \in \mathbb{Z}_+\}$ (with $\rho_0 = 0$). Note that we do not precise the dependence on N in the notation. In our case, the probability that the coupling is successful

$$p_N := \inf_{\bar{n} \in [N]^{m_N}} \mathbb{P}_{\bar{n}}((\underline{R}_{\rho_1}^{N,1}, \ldots, \underline{R}_{\rho_1}^{N,m_N}) = (R_{\rho_1}^{N,1}, \ldots, R_{\rho_1}^{N,m_N}))$$

$$= 1 - \sup_{\bar{n} \in [N]^{m_N}} \mathbb{P}_{\bar{n}}((\underline{R}_{\rho_1}^{N,1}, \ldots, \underline{R}_{\rho_1}^{N,m_N}) \neq (R_{\rho_1}^{N,1}, \ldots, R_{\rho_1}^{N,m_N}))$$

$$= 1 - \sup_{\bar{n} \in [N]^{m_N}} \|\mathbb{P}_{\bar{n}}((R_{\rho_1}^{N,1}, \ldots, R_{\rho_1}^{N,m_N}) = \cdot) - (\nu^N)^{\otimes m_N}(\cdot)\|_{TV}$$

where $\mathbb{P}_{\bar{n}}$ stands for the law of $\{R_g^{N,1}, \ldots, R_g^{N,m_N}\}_{g \geq 0}$ (or $\{\underline{R}_g^{N,1}, \ldots, \underline{R}_g^{N,m_N}\}_{g \geq 0}$) starting at the state $\bar{n} \in [N]^{m_N}$ and where Proposition 4.7 in Levin and Peres (2017) is used for the last equality. To prove that $p_N \to 1$ when $N \to \infty$, take $\varepsilon > 0$ such that $N^\varepsilon \tau_N c_N \to 0$. The condition $\mu^N(1) > 0$ implies that, for any $i \geq 1$, the processes $\{R_g^{N,i}\}_{g \geq 0}$ are irreducible. So, by equation (4.29) in Levin and Peres (2017) and the definition of τ_N, we have

$$\|\mathbb{P}_{\bar{n}}((R_{N^\varepsilon \tau_N}^{N,1}, \ldots, R_{N^\varepsilon \tau_N}^{N,m_N}) = \cdot) - (\nu^N)^{\otimes m_N}(\cdot)\|_{TV} < (1/4)^{N^\varepsilon}. \tag{2.10}$$

Then observe that, stochastically, $\rho_1 \geq \Gamma^N$ where Γ^N is a geometric random variable with parameter $n^2 c_N \geq \binom{n}{2} c_N$ (think of a Cannings model without dormancy) and thus

$$\mathbb{P}(\rho_1 < N^\varepsilon \tau_N) \leq \mathbb{P}(\Gamma^N < N^\varepsilon \tau_N) = 1 - (1 - n^2 c_N)^{N^\varepsilon \tau_N} \to 0.$$

This implies that $p_N \geq 1 - (1/4)^{N^\varepsilon}(1 + o(1))$.

Let $T_1^N = \inf\{i \geq 1 : |\bar{Z}_{\rho_i}^N| = 1\}$. Observe that if $|\bar{Z}_0^N| = n$ (we will use the notation $\mathbb{P}_n$), stochastically

$$T_1^N \leq \sum_{i=1}^n G_i^N \tag{2.11}$$

where the G_i^N's are independent geometric random variables with parameter β_N^{-2} (the probability that two particles in stationarity reach state 1). We finish the proof of (2.4) noting that the trajectories of both processes are identical with overwhelming probability

$$\mathbb{P}_n(\sup_t \||\bar{Z}_{\lfloor t\beta_N^2/c_N \rfloor}^N| - |\bar{B}_{\lfloor t\beta_N^2/c_N \rfloor}^N|\| = 0) = \mathbb{E}[p_N^{T_1^N}] \geq \mathbb{E}[p_N^{\sum_{i=1}^n G_i^N}]$$

$$= \left(\frac{p_N \beta_N^{-2}}{1 - (1 - \beta_N^{-2})p_N} \right)^n$$

$$= \left(1 - \beta_N^2 + p_N^{-1} \beta_N^2 \right)^{-n}.$$

Since $p_N \geq 1 - (1/4)^{N^\varepsilon}(1 + o(1))$ and $(1/4)^{N^\varepsilon} \beta_N^2 \to 0$, we have that the quantity above converges to 1 when $N \to \infty$.

Finally, let us prove (2.5). Let $t > 0$ fixed, and suppose that ν^N converges to a measure ν. By, equation (2.9) we have that $\lim_{N \to \infty} |\bar{Z}_{\lfloor t\beta_N^2/c_N \rfloor}^N| = N_t^K$. Then, observe that $\bar{Z}_g^N$ has a multinomial distribution with parameters $|\bar{Z}_g^N|$ and ν^N. Thus, in distribution,

$$\lim_{N \to \infty} \bar{Z}_{\lfloor t\beta_N^2/c_N \rfloor}^N = V^{t,K}. \tag{2.12}$$

On the other hand, by (2.4), we have that

$$\lim_{N \to \infty} |\bar{B}_{\lfloor t\beta_N^2/c_N \rfloor}^N| = N_t^K.$$

For $t > 0$ fixed, let us couple $\bar{Z}^N_{\lfloor t\beta^2_N/c_N \rfloor}$ and $\bar{B}^N_{\lfloor t\beta^2_N/c_N \rfloor}$ to show that the limit (2.12) is the same for $\bar{B}^N_{\lfloor t\beta^2_N/c_N \rfloor}$. As we did before, the coupling consists in constructing the random variable $(\underline{R}^{N,1}_{\lfloor t\beta^2_N/c_N \rfloor}, \ldots, \underline{R}^{N,m_N}_{\lfloor t\beta^2_N/c_N \rfloor})$ as the optimal coupling of $(R^{N,1}_{\lfloor t\beta^2_N/c_N \rfloor}, \ldots, R^{N,m_N}_{\lfloor t\beta^2_N/c_N \rfloor})$, which depends on the initial condition $\bar{n} \in [N]^{m_N}$. The probability that the coupling is successful is

$$\varrho_N := \inf_{\bar{n} \in [N]^{m_N}} \mathbb{P}_{\bar{n}}((\underline{R}^{N,1}_{\lfloor t\beta^2_N/c_N \rfloor}, \ldots, \underline{R}^{N,m_N}_{\lfloor t\beta^2_N/c_N \rfloor}) = (R^{N,1}_{\lfloor t\beta^2_N/c_N \rfloor}, \ldots, R^{N,m_N}_{\lfloor t\beta^2_N/c_N \rfloor}))$$

$$= 1 - \sup_{\bar{n} \in [N]^{m_N}} \mathbb{P}_{\bar{n}}((\underline{R}^{N,1}_{\lfloor t\beta^2_N/c_N \rfloor}, \ldots, \underline{R}^{N,m_N}_{\lfloor t\beta^2_N/c_N \rfloor}) \neq (R^{N,1}_{\lfloor t\beta^2_N/c_N \rfloor}, \ldots, R^{N,m_N}_{\lfloor t\beta^2_N/c_N \rfloor}))$$

$$= 1 - \sup_{\bar{n} \in [N]^{m_N}} \|\mathbb{P}_{\bar{n}}((R^{N,1}_{\lfloor t\beta^2_N/c_N \rfloor}, \ldots, R^{N,m_N}_{\lfloor t\beta^2_N/c_N \rfloor}) = \cdot) - (\nu^N)^{\otimes m_N}(\cdot)\|_{TV}$$

Take $\varepsilon > 0$ such that $N^\varepsilon \tau_N c_N \to 0$, and observe that $\beta_N > 0$. This implies that

$$\mathbb{P}(\lfloor t\beta^2_N/c_N \rfloor < N^\varepsilon \tau_N) \to 0 \text{ as } N \to \infty. \tag{2.13}$$

By, (2.10) and (2.13), we have that $\varrho_N \to 1$. This gives (2.5). $\qquad\square$

Remark 2.10. Consider two processes, $\{\underline{R}^N_g\}_{g \geq 0}$ and $\{R^N_g\}_{g \geq 0}$, the first one starting with one particle in stationarity and the second one starting with one particle in state one. If we consider their Doebling coupling (which consists in letting them evolve according to their respective laws and in merging their paths when they meet for the first time, see Levin and Peres (2017), Chapter 5), their coupling time, T^N, is less than two with probability

$$\nu^N(1) + \sum_{i=1}^{m_N-1} \mu^N(i)\nu^N(i+1) \leq \nu^N(1) + \frac{1}{\beta_N} \sum_{i=1}^{m_N-1} \mu^N(i) = \frac{1}{\beta_N}(2 - \mu^N(m_N)).$$

As the process $\{R^N_g\}_{g \geq 0}$ visits the state one approximately every β_N steps we conclude that $\mathbb{P}(T^N > N^\varepsilon \beta^2_N) \to 0$ when $N \to \infty$. Since $\tau_N \leq \inf\{t \geq 0; \mathbb{P}(T^N > t) < 1/4\}$, we obtain that $m_N \leq \tau_N \leq N^\varepsilon \beta^2_N$ for N large enough. Note that m_N is always a lower bound for τ_N since the mixing time is at least the time for one particle starting at m_N to reach the origin. Then hypotheses of Theorem 2.6 can be relaxed to the following

$$c_N/\beta^2_N \to 0, \quad m_N/(N^\varepsilon \beta^2_N) \to \delta < 1, \quad N^{2\varepsilon}\beta^2_N c_N \to 0, \quad (1/4)^{N^\varepsilon}\beta^2_N \to 0 \quad \text{and} \quad d_N/(\beta_N c_N) \to 0$$

with the advantage that they are easier to verify.

Proof of Theorem 2.8: The proof of (2.7) is similar to that of (2.4) in Theorem 2.6. First observe that (2.6) implies that $c_N \to 0$. Assuming furthermore that $\beta_N \to \beta < \infty$, we get that $c_N/\beta^2_N \to 0$, $(1/4)^{N^\varepsilon}\beta^2_N \to 0$ and $N^\varepsilon \tau_N c_N \to 0$ (since the mixing time is of order 1 in this case). These are the three first conditions of (2.3), necessary to mimic the proof of Theorem 2.6.

In the present case, let I_i denote the indicator of the event that $L^{N,i}_1 = L^{N,j}_1$ for some $j \in [i-1]$. Note that $\{C^N_{\lfloor t/c_N \rfloor}\}_{t \geq 0}$ has generator

$$\mathcal{C}^N f(n) = c_N^{-1} \mathbb{E}[f(n - \sum_{i=1}^n I_i) - f(n)]$$

which by hypothesis converges to the generator of the block counting process of a Ξ-coalescent. Finally note that, using the same notation, the generator of the artificial (in stationarity) block counting process $\{\bar{Z}^N_{\lfloor t/c_N \rfloor}\}_{t \geq 0}$ is

$$\mathcal{C}^N f(n) = c_N^{-1} \mathbb{E}[f(n - \sum_{i=1}^n I_i \mathbf{1}_{\{\underline{R}^{N,i}_1 = 1\}}) - f(n)].$$

As $1_{\{\underline{R}_1^{N,i}=1\}}$ is a Bernoulli random variable with parameter tending to β^{-1} and independent of I_i, we conclude that

$$\lim_{N\to\infty}\{|\bar{B}^N_{\lfloor t/c_N\rfloor}|\}_{t\geq 0}=\{N_t^{\Xi^\beta}\}_{t\geq 0}.$$

The rest of the proof is identical to that of Theorem 2.6. □

Remark 2.11. The window process can also be defined in terms of a modification of the partition-valued ancestral process. This more formal (and more complicated) definition follows the historical approach of Kaj et al. (2001); Blath et al. (2013). Fix a generation g_0, and $S_{g_0}\subset\cup_{i=1}^{m_N}V^N_{g_0+1-i}$. For $g\geq 1$, consider the equivalence relation on S_{g_0}, that we denote by $\sim_g$, such that $u\sim_g v$ if and only if they have a common ancestor at a generation between g_0-g+1 and g_0. Let $\pi_0=\pi_1$ be the trivial partition made of the isolated elements of S_{g_0} (singletons) and let π_g be the partition induced by $\sim_g$ in the sample S_{g_0}. Let $\mathcal{B}^N_g$ be the set composed by the closest ancestors, living at a generation g_0-g' for some $g'\geq g$, of each of the blocks in π_g. Then, for $1\leq i\leq m_N$, we define

$$B^{N,i}_g:=|\mathcal{B}^N_g\cap V^N_{g_0-g+1-i}|.$$

We illustrate this definition by the realization pictured in Figure 2.1.

In this case $\pi_0=\pi_1=\pi_2=\{\{v_1\},\{v_2\},\{v_3\},\{v_4\},\{v_5\}\}$. Observe that even if some individuals reach their common ancestor at generation -2, they remain isolated in π_2. Then $\pi_3=\{\{v_1,v_2\},\{v_3,v_4\},\{v_5\}\}$, $\pi_4=\{\{v_1,v_2,v_3,v_4\},\{v_5\}\}$, $\pi_5=\{\{v_1,v_2,v_3,v_4,v_5\}\}$. Hence, $\mathcal{B}^8_0=\{v_1,v_2,v_3,v_4,v_5\}$ and, when moving some generations backwards, we get $\mathcal{B}^8_1=\{v_5,(-1,5),(-2,4),(-2,7),(-2,7)\}$ and $\mathcal{B}^8_2=\{(-2,3),(-2,4),(-2,4),(-2,7),(-2,7)\}$. Observe that in $\mathcal{B}^8_2$ the ancestors $(-2,4)$ and $(-2,7)$ appear twice. Also $\mathcal{B}^8_3=\{(-3,7),(-3,7),(-4,6)\}$, $\mathcal{B}^8_4=\{(-4,6),(-4,6)\}$, $\mathcal{B}^8_5=\{(-5,5)\}$.

Finally, note that the coupled variables $Y^{N,j}_g=(R^{N,j}_g,L^{N,j}_g)$ define the process that models the distance between g and the position of the ancestor of the j-th block induced by $\sim_g$.

3. The forward frequency process

In this section we introduce the forward frequency process associated to the seed bank graph, we establish duality results with the ancestral and window processes introduced in the previous section, and we establish some scaling limits results thanks to these tools.

Definition 3.1 (The frequency process). Fix a generation g_0 and an initial sample

$$S_{g_0}\subset\cup_{i=1}^{m_N}V^N_{g_0+1-i},$$

that we call the type A individuals. Hence, $\cup_{i=1}^{m_N}V^N_{g_0+1-i}\backslash S_{g_0}$ is the set of type a individuals. For $g\geq 0$, set (omitting again the dependence to S_{g_0})

$$X^{N,i}_g=\frac{1}{N}|\{v\in V^N_{g_0+g+1-i}:v\text{ is not connected to }u\text{ for some }u\in\cup_{i=1}^{m_N}V^N_{g_0+1-i}\backslash S_{g_0}\}|.$$

Then, define the process of the neutral frequency of type A individuals $\{\bar{X}^N_g\}_{g\geq 0}$, by

$$\bar{X}^N_g=(X^{N,1}_g,\ldots,X^{N,m_N}_g).$$

Set a vector $\bar{x}=(x_1,\ldots,x_{m_N})\in([N]/N)^{m_N}$. In the sequel, we suppose that the forward frequency process starts from a fraction x_1 of generation 0, a fraction x_2 of generation -1, and so on... We denote this sample by $S_0(\bar{x})=\cup_{i=1}^{m_N}\cup_{k=1}^{x_iN}\{(1-i,k)\}$ and we denote the law of the frequency process starting from this configuration by $\mathbf{P}_{\bar{x}}$.

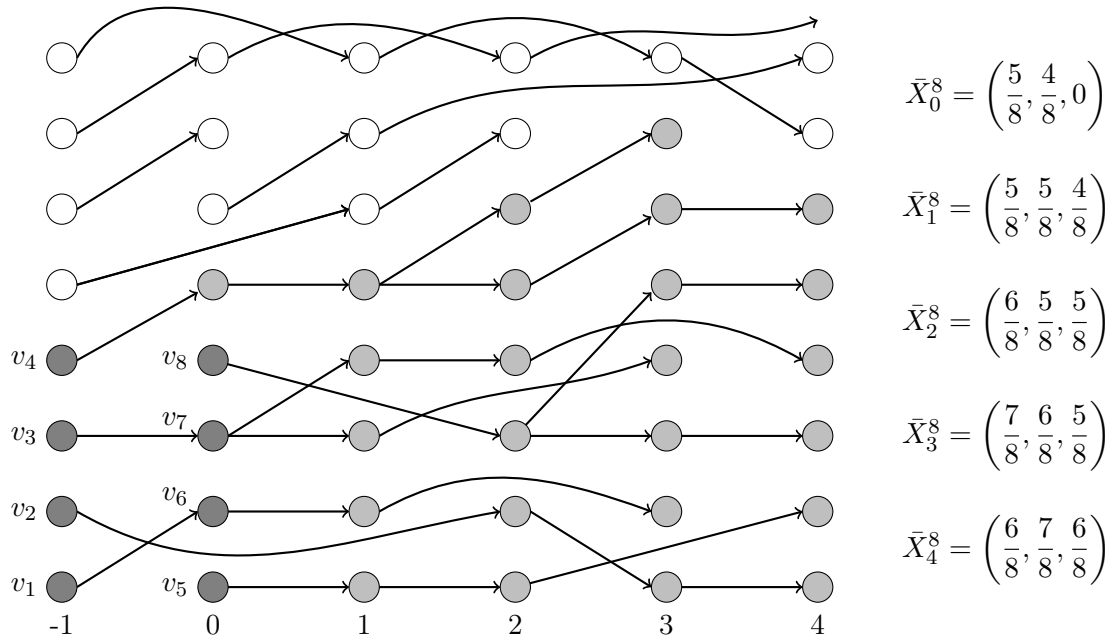Again for simplicity, we suppose that $\sup\{i\geq 1:x_i>0\}$ does not depend on N.

$$\bar{X}_0^8 = \left(\frac{5}{8}, \frac{4}{8}, 0\right)$$

$$\bar{X}_1^8 = \left(\frac{5}{8}, \frac{5}{8}, \frac{4}{8}\right)$$

$$\bar{X}_2^8 = \left(\frac{6}{8}, \frac{5}{8}, \frac{5}{8}\right)$$

$$\bar{X}_3^8 = \left(\frac{7}{8}, \frac{6}{8}, \frac{5}{8}\right)$$

$$\bar{X}_4^8 = \left(\frac{6}{8}, \frac{7}{8}, \frac{6}{8}\right)$$

FIGURE 3.3. In this case $N = 8$, $m_N = 3$ and $\bar{x} = \left(\frac{4}{8}, \frac{4}{8}, 0\right)$. The gray circles represent the members of $S_0(\bar{x}) = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$ where, for example, $v_3 = (-1, 3)$ and $v_6 = (0, 2)$. The light gray circles represent the sample's offspring. It is useful to observe that $X_g^{N,i} = X_{g+1}^{N,i+1}$ for $i < m_N$.

Proposition 3.2. *Fix the parameters N, μ^N and $\mathcal{W}^N$ of the seed bank di-graph. The processes $\{\bar{X}_g^N\}_{g \geq 0}$ and $\{\bar{A}_g^N\}_{g \geq 0}$ are sampling duals: for every $g \geq 0$, we have*

$$\mathbf{E}_{\bar{x}}[h^0(\bar{n}, \bar{X}_g^N)] = \mathbb{E}_{\bar{n}}[h^0(\bar{A}_g^N, \bar{x})]$$

where $h^0(\bar{n}, \bar{x}) := \mathbb{P}_{\bar{n}}(\mathcal{A}_1^N \subset S_0(\bar{x}))$.

Proof: Suppose that the ancestral process starts at generation $g + 1$ from the sample $S_{g+1}(\bar{n})$, as in Definition 2.2. Also suppose that the frequency process starts at generation 0 from the sample $S_0(\bar{x})$, as in Definition 3.1. Introduce the functions

$$h^g(\bar{n}, \bar{x}) := \mathbb{P}_{\bar{n}}(\mathcal{A}_{g+1}^N \subset S_0(\bar{x})). \tag{3.1}$$

We can write $h^g(\bar{n}, \bar{x})$ in terms of the forward process by conditioning as follows.

$$h^g(\bar{n}, \bar{x}) = \sum_{\bar{y} \in ([N]/N)^{m_N}} h^0(\bar{n}, \bar{y}) \mathbf{P}_{\bar{x}}(\bar{X}_g^N = \bar{y})$$

$$= \mathbf{E}_{\bar{x}}[h^0(\bar{n}, \bar{X}_g^N)].$$

At this point it should be clear that we can also condition according to the backward process.

$$h^g(\bar{n}, \bar{x}) = \sum_{\bar{m} \in [N]^{m_N}} h^0(\bar{m}, \bar{x}) \mathbb{P}_{\bar{n}}(\bar{A}_g^N = \bar{m})$$

$$= \mathbb{E}_{\bar{n}}[h^0(\bar{A}_g^N, \bar{x})].$$

This implies that for all $\bar{x} \in ([N]/N)^{m_N}$, $\bar{n} \in [N]^{m_N}$ and $g \geq 1$,

$$\mathbb{E}_{\bar{n}}[h^0(\bar{A}_g^N, \bar{x})] = \mathbf{E}_{\bar{x}}[h^0(\bar{n}, \bar{X}_g^N)].$$

$\square$

The sampling duality provides a relation between the forward and the ancestral processes. However, in this case the relation is not a moment duality. It is possible to write explicitly this sampling duality and to use it, but we will rather use the less natural window process which has the advantage of being precisely the moment dual of the forward process.

Proposition 3.3. *Fix the parameters N, μ^N and $\mathcal{W}^N$ of the seed bank di-graph. The window process $\{\bar{B}_g^N\}_{g\geq 0}$ and the forward frequency process $\{\bar{X}_g^N\}_{g\geq 0}$ are moment duals.*

Proof: We will construct a sampling duality that is exactly moment duality i.e duality with respect to the function $H : \mathbb{Z}_+^{m_N} \times [0,1]^{m_N} \mapsto [0,1]$,

$$H(\bar{n}, \bar{x}) = \prod_{i=1}^{m_N} x_i^{n_i} \tag{3.2}$$

Fix $\bar{n} \in [N]^{m_N}$ and $\bar{x} \in ([N]/N)^{m_N}$, and set the samples $S_0(\bar{n})$ and $S_0(\bar{x})$ as in Definition 2.2 and Definition 3.1 (with $g_0 = 0$). Observe that $S_0(\bar{n}) = \{v = (1 - i, U_{j,i}), i = 1, \ldots, m_N, j = 1, \ldots, n_i\}$ where the $U_{j,i}$'s form a family of independent uniformly distributed random variables with values in $[N]$. Then, we have

$$\tilde{h}(\bar{n}, \bar{x}) := \mathbb{P}(S_0(\bar{n}) \subset S_0(\bar{x})) = \prod_{i=1}^{m_N}\prod_{j=1}^{n_i} \mathbb{P}((1 - i, U_{j,i}) \in S_0(\bar{x})) = \prod_{i=1}^{m_N}\prod_{j=1}^{n_i} x_i = H(\bar{n}, \bar{x}).$$

Now we prove sampling duality with respect to this function. As in the proof of Proposition 3.2, condition on $\bar{X}_g^N$ to obtain that $\mathbb{P}_{\bar{n}}(\mathcal{B}_g^N \subset S_0(\bar{x})) = \mathbf{E}_{\bar{x}}[\tilde{h}(\bar{n}, \bar{X}_g^N)]$ and condition on $\bar{B}_g^N$ to obtain that $\mathbb{P}_{\bar{n}}(\mathcal{B}_g^N \subset S_0(\bar{x})) = \mathbb{E}_{\bar{n}}[\tilde{h}(\bar{B}_g^N, \bar{x})]$. $\square$

Now we are able to state an analogue of Theorem 2.6 for the dual process, using the moment duality.

Theorem 3.4 (Convergence of the forward frequency process)**.** *Assume that $m_N \leq m < \infty$ for all $N \in \mathbb{Z}_+$. Fix $\{\mathcal{W}^N\}_{N\geq 1}$ and $\{\mu^N\}_{N\geq 1}$ (and the associated stationary distribution ν^N) such that either the assumptions of Theorem 2.6 hold or the assumptions of Theorem 2.8 hold. Suppose that ν^N converges to a measure ν on $[m]$ as $N \to \infty$. Let $\{\bar{X}^N\}_{N\geq 1}$ be the sequence of frequency processes with parameters N, $\mathcal{W}^N$ and μ^N and starting condition $\bar{X}_0^N = (\lfloor Nx_1 \rfloor/N, \ldots, \lfloor Nx_m \rfloor/N)$ for some $\bar{x} \in [0,1]^m$.*
i) Under the assumptions of Theorem 2.6,

$$\lim_{N\to\infty} \{\bar{X}_{\lfloor t\beta_N^2/c_N \rfloor}^N\}_{t\geq 0} = \{\bar{X}_t\}_{t\geq 0}$$

in the finite dimension sense, where $\bar{X}_t$ is a vector with m identical coordinates X_t such that $X_0 = x_0 = \sum_{i=1}^m \nu(i)x_i$ a.s., and $\{X_t\}_{t\geq 0}$ is the Wright-Fisher diffusion (dual of $\{N_t^K\}_{t\geq 0}$).
ii) Under the assumptions of Theorem 2.8,

$$\lim_{N\to\infty} \{\bar{X}_{\lfloor t/c_N \rfloor}^N\}_{t\geq 0} = \{\bar{X}_t\}_{t\geq 0}$$

in the finite dimension sense, where $\bar{X}_t$ is a vector with m identical coordinates X_t such that $X_0 = x_0 = \sum_{i=1}^m \nu(i)x_i$ a.s., and $\{X_t\}_{t\geq 0}$ is the moment dual of $\{N_t^{\Xi^\beta}\}_{t\geq 0}$.

Remark 3.5. It is surprising at first that the components of the limit are identical. However, this becomes intuitive when observing that the support of μ^N vanishes on the limiting time scale. Theorem 3.4 uses the more restrictive assumption that the support for $\{\mu^N\}_{N\geq 1}$ is bounded. This seems to be more than a technical assumption, because it is hard to believe that an asymptotically infinite dimensional sequence of processes would converge to an infinite dimensional process with all entries being equal (even if the support of μ^N vanishes on the limiting time scale). A natural

question is, what is the limit in this more general scenario? An interesting framework to study this question, in which there are results in the literature, is taking $\mu^N := \mu$ such that $\mu(\{j : j > i\}) \sim i^{-\alpha}$ see Blath et al. (2013). If $\alpha > 1$, Theorem 3.4 should essentially apply, but the convergence should hold only for finitely (but arbitrarily big) many dimensions simultaneously. If $\alpha \in (0, 1/2)$ the limit, if it exists, should be related to fractional Brownian motion (see Hammond and Sheffield (2013) and Igelbrink and Wakolbinger, 2023). The frequency process in the case $\alpha \in (0, 1/2)$ is believed to exist and was named by Blath and Spanò the Fractional Wright Fisher Diffusion. If one takes $\mu^N = 1/N\delta_1 + Geo(1/N)$ a modification of Theorem 3.4 leads to a criterion for convergence to Ξ-seed bank diffusions, this is current work of the authors. The seed bank diffusion was introduced in Blath et al. (2016) and can be thought of as a delayed stochastic differential equation Blath et al. (2019) (see also Blath et al., 2023).

Proof: We only write the details for case *i)*, case *ii)* follows identically. The proof is a consequence of Proposition 3.3, Theorem 2.6 and the moment problem. Let us abuse the notation and write $\bar{X}_0^N = \bar{x}$ for every N.

First, let us clarify the role of x_0. Recall that the process $\{X_t\}_{t \geq 0}$ is a martingale. In particular, its expectation remains constant. We claim that for every $i \in [m]$, $\lim_{N \to \infty} \mathbf{E}_{\bar{x}}[X_{\lfloor t\beta_N^2/c_N \rfloor}^{N,i}] = x_0$. To see this we use duality and convergence to stationarity of a single dual particle.

$$\lim_{N \to \infty} \mathbf{E}_{\bar{x}}[X_{\lfloor t\beta_N^2/c_N \rfloor}^{N,i}] = \lim_{N \to \infty} \mathbb{E}_{e_i}\left[\prod_{j=1}^m x_j^{B_{\lfloor t\beta_N^2/c_N \rfloor}^{N,j}}\right] = \sum_{i=1}^m x_i \nu(i) = x_0. \tag{3.3}$$

The first equality comes from duality. For the second equality, recall that in both Theorem 2.6 and Theorem 2.8 we suppose that $\beta_N^2/c_N \to \infty$ and thus that the process is in stationarity in the limit. The third equality follows from the fact that there is only one positive entry of the unitary vector $\bar{B}_{\lfloor t\beta_N^2/c_N \rfloor}^N$ and that the position of the entry with the one is ν-distributed in the limit.

Now let us study the limiting behavior of one coordinate. Let $n \geq 1$.

$$\begin{aligned}
\lim_{N \to \infty} \mathbf{E}_{\bar{x}}[(X_{\lfloor t\beta_N^2/c_N \rfloor}^{N,1})^n] &= \lim_{N \to \infty} \mathbf{E}_{\bar{x}}[H(n.e_1, \bar{X}_{\lfloor t\beta_N^2/c_N \rfloor}^N)] \\
&= \lim_{N \to \infty} \mathbb{E}_{n.e_1}[H(\bar{B}_{\lfloor t\beta_N^2/c_N \rfloor}^N, \bar{x})] \\
&= \mathbb{E}_{n.e_1}[H(V^{t,K}, \bar{x})] \\
&= \mathbb{E}_n[x_0^{N_t^K}] \\
&= \mathbf{E}_{x_0}[X_t^n].
\end{aligned}$$

The third equality follows from (2.5) and in the fourth equality we used the same argument as for (3.3). This proves that all the moments of $X_{\lfloor t\beta_N^2/c_N \rfloor}^{N,1}$ converge to the moments of the Wright-Fisher diffusion.

Finally, we check that in the limit all the coordinates of $\bar{X}_{\lfloor t\beta_N^2/c_N \rfloor}^N$ must take the same value. To do this we will calculate the square of the difference of two arbitrary coordinates. Let $i, j \in [m]$.

With the same arguments as before,

$$
\begin{aligned}
\lim_{N \to \infty} \mathbf{E}_{\bar{x}}[(X^{N,i}_{\lfloor t\beta_N^2/c_N \rfloor} - X^{N,j}_{\lfloor t\beta_N^2/c_N \rfloor})^2] &= \lim_{N \to \infty} \Big[\mathbf{E}_{\bar{x}}[(X^{N,i}_{\lfloor t\beta_N^2/c_N \rfloor})^2] + \mathbf{E}_{\bar{x}}[(X^{N,j}_{\lfloor t\beta_N^2/c_N \rfloor})^2] \\
&\quad -2\mathbf{E}_{\bar{x}}[X^{N,i}_{\lfloor t\beta_N^2/c_N \rfloor} X^{N,j}_{\lfloor t\beta_N^2/c_N \rfloor}] \Big] \\
&= \lim_{N \to \infty} \Big[\mathbf{E}_{\bar{x}}[H(2e_i, \bar{X}^N_{\lfloor t\beta_N^2/c_N \rfloor})] + \mathbf{E}_{\bar{x}}[H(2e_j, \bar{X}^N_{\lfloor t\beta_N^2/c_N \rfloor})] \\
&\quad -2\mathbf{E}_{\bar{x}}[H(e_i + e_j, \bar{X}^N_{\lfloor t\beta_N^2/c_N \rfloor})] \Big] \\
&= \lim_{N \to \infty} \Big[\mathbb{E}_{2e_i}[H(\bar{B}^N_{\lfloor t\beta_N^2/c_N \rfloor}, \bar{x})] + \mathbb{E}_{2e_j}[H(\bar{B}^N_{\lfloor t\beta_N^2/c_N \rfloor}, \bar{x})] \\
&\quad -2\mathbb{E}_{e_i+e_j}[H(\bar{B}^N_{\lfloor t\beta_N^2/c_N \rfloor}, \bar{x})] \Big] \\
&= \mathbb{E}_{2e_i}[H(V^{t,K}, \bar{x})] + \mathbb{E}_{2e_j}[H(V^{t,K}, \bar{x})] \\
&\quad - 2\mathbb{E}_{e_i+e_j}[H(V^{t,K}, \bar{x})] \\
&= 0.
\end{aligned}
$$

This ends the proof.

$\square$

4. Extended seed bank di-graph with mutations

It shall be interesting to study a variant of the model where mutations are added. There is a classical duality relation between the Wright-Fisher diffusion with mutations

$$
dX_t = (u_1(1 - X_t) - u_2 X_t)dt + \sqrt{X_t(1 - X_t)}dB_t. \tag{4.1}
$$

and the block counting process of a Kingman coalescent with freezing, where every lineage can disappear at rate $u_1 + u_2$. This relation was established in Etheridge and Griffiths (2009) (see also Griffiths and Spanó (2010) and the seminal work of Ethier and Griffiths, 1993) and was generalized in Etheridge et al. (2010) to Λ-coalescents with freezing.

These works motivate that we modify the seed bank di-graph to include mutations (Definition 4.1). In order to observe genealogies with freezing in the model we consider that mutations come from a separate source and not from a reproduction event. This will let us establish a duality relation even in the finite population case (Proposition 4.5) between a modification of the window process and a modification of the forward frequency process, and hence, a duality relation in the limit generalizing known results (Theorem 4.10). The limit genealogical processes obtained in this relation are described in Theorems 4.7 and 4.9. They are general coalescents but with a freezing component. We give a formal description of them that is more convenient for our setting than the common definition in Definition 4.6. Most of the proofs of this section are very similar to those of Sections 2 and 3. We emphasize the modifications and the intuitions in the proof of Theorem 4.7 and enunciate the other results without the proofs, leaving them to the reader. For sake of simplicity, most of the notations of this section will be identical to those of the previous ones when the objects describe the same concepts, although their definitions are slightly modified.

Definition 4.1 (The extended seed bank random di-graph). Set N, μ^N and $\mathcal{W}^N$ as in Definition 2.1, the vertex set V^N, and also the random variables $\{\bar{W}^N_g\}_{g \in \mathbb{Z}}$, where $\bar{W}^N_g = \{W^N_v\}_{v \in V^N_g}$, $\{J^N_v\}_{v \in V^N}$ and $\{U^N_v\}_{v \in V^N}$. Fix two more parameters $u_{N,1}, u_{N,2} \geq 0$ such that $u_{N,0} := 1 - u_{N,1} - u_{N,2} \geq 0$, and let $\{K^N_v\}_{v \in V^N}$ be a sequence of independent random variables with state space $\{0, 1, 2\}$ such that $\mathbb{P}(K^N_v = i) = u_{N,i}$. Consider the extended vertex set $V^N \cup \{\Delta_1\} \cup \{\Delta_2\}$ and the random function

$F : V^N \mapsto V^N \cup \{\Delta_1\} \cup \{\Delta_2\}$ defined by the rule

$$F(v) = \begin{cases} (g(v) - J_v^N, U_v^N) & \text{if } K_v^N = 0 \\ \Delta_i & \text{if } K_v^N = i \in \{1, 2\}. \end{cases} \tag{4.2}$$

Let E^N be the set of directed edges

$$E^N = \{\{v, F(v)\}, \text{ for all } v \in V^N\}.$$

The extended seed bank random di-graph with parameters N, μ^N, $\mathcal{W}^N$ and $u_{N,1}, u_{N,2}$ is given by $G^N := (V^N \cup \{\Delta_1\} \cup \{\Delta_2\}, E^N)$.

In this extended version of the graph, we can define a modification of the original window process.

Definition 4.2 (The window process with mutations)**.** Fix a generation g_0, and $S_{g_0} \subset \cup_{i=1}^{m_N} V_{g_0+1-i}^N$. The window process with mutations is the chain $\{\bar{B}_g^N, D_g^N\}_{g \geq 0}$ where $\{\bar{B}_g^N\}_{g \geq 0}$ is the window process introduced in Definition 2.4 (but associated to the extended di-graph) and $\{D_g^N\}_{g \geq 0}$ is the process counting the cumulate number of lineages connecting with Δ_1 or Δ_2 under the rule F, that is

$$D_g^N = |\{(v, v') \in E^N, g(v) \geq g_0 - g, v \text{ is an ancestor of some individual of } S_{g_0} \text{ and } v' \in \{\Delta_1, \Delta_2\}\}|$$

We denote by $\mathbb{P}_{\bar{n}}$ the law of the window process starting from $S_{g_0}(\bar{n})$ and with $D_0^N = 0$.

Example 4.3. Modify Figure 2.1 such that individual $(-2, 7)$ is connected to Δ_1. The window process with mutations has the following values: $\{\bar{B}_0^8, D_0^8\} = \{(4, 1), 0\}$, $\{\bar{B}_1^8, D_1^8\} = \{(2, 3), 0\}$, $\{\bar{B}_2^8, D_2^8\} = \{(5, 0), 0\}$, $\{\bar{B}_3^8, D_3^8\} = \{(1, 1), 1\}$, $\{\bar{B}_4^8, D_4^8\} = \{(2, 0), 1\}$, $\{\bar{B}_5^8, D_5^8\} = \{(1, 0), 1\}$.

The state Δ_1 can be seen as the source of type a mutations and the state Δ_2 as the source of type A mutations. So we can slightly modify Definition 3.1 to obtain the new forward frequency process.

Definition 4.4 (The frequency process with mutations)**.** Fix a generation g_0 and an initial sample $S_{g_0} \subset \cup_{i=1}^{m_N} V_{g_0+1-i}^N$, that we call the type A individuals. Hence, $\cup_{i=1}^{m_N} V_{g_0+1-i}^N \backslash S_{g_0}$ is the set of type a individuals. For $g \geq 0$, set (omitting again the dependence to S_{g_0})

$$X_g^{N,i} = \frac{1}{N} |\{v \in V_{g_0+g+1-i}^N : v \text{ is not connected to } \Delta_1 \text{ nor to } u \text{ for some } u \in \cup_{i=1}^{m_N} V_{g_0+1-i}^N \backslash S_{g_0}\}|.$$

Then, define the process of the neutral frequency of type A individuals $\{\bar{X}_g^N, \theta^N\}_{g \geq 0}$, where $\theta^N = u_{N,1}/(u_{N,1} + u_{N,2})$ and

$$\bar{X}_g^N = (X_g^{N,1}, \dots, X_g^{N,m_N}).$$

Set a vector $\bar{x} = (x_1, \dots, x_{m_N}) \in ([N]/N)^{m_N}$. In the sequel, we suppose that the forward frequency process starts from a fraction x_1 of generation 0, a fraction x_2 of generation -1, and so on. We denote this sample by $S_0(\bar{x}) = \cup_{i=1}^{m_N} \cup_{k=1}^{x_i N} \{(1-i, k)\}$ and we denote the law of the frequency process starting from this configuration by $\mathbf{P}_{\bar{x}}$.

We obtain a duality result, which is the analogue of the moment duality obtained in Proposition 3.3, and that can easily be proved adapting the proofs of Section 3. It is inspired by Etheridge and Griffiths (2009) and Griffiths and Spanó (2010).

Proposition 4.5. *Fix N, μ^N, $\mathcal{W}^N$ and $u_{N,1}, u_{N,2}$. The window process $\{\bar{B}_g^N, \bar{D}_g^N\}_{g \geq 0}$ and the frequency process with mutations $\{\bar{X}_g^N, \theta^N\}_{g \geq 0}$ are moment duals, in the sense that for any $g \geq 0$,*

$$\mathbf{E}_{\bar{x}} \left[\prod_{i=1}^{m_N} (X_g^{N,i})^{n_i} \right] = \mathbb{E}_{\bar{n}} \left[(\theta^N)^{D_g^N} \prod_{i=1}^{m_N} x_i^{B_g^{N,i}} \right].$$

Let us recall the concept of coalescent with freezing. Usually this model describes an exchangeable coalescent where, apart from participating to merging events, lineages can disappear (or freeze) at a constant rate, independent of each other. It appears in general population models with mutations in Etheridge et al. (2010). Its Kingman version is also crucial in the infinite alleles model, where Ewens' sampling formula is established Ewens (1972), but also to provide asymptotics to the seed bank (or peripatric) coalescent González Casanova et al. (2022b).

As could be anticipated from Definition 4.2, we will modify a bit this object keeping track of the whole genealogy and of the frozen lineages separately. We only need to consider its block counting process for our purpose

Definition 4.6 (The block counting process of a coalescent with freezing). The block counting process of a Ξ-coalescent with freezing parameter u is the process $\{M_t^{\Xi}, D_t\}_{t \geq 0}$ where M_t^{Ξ} counts the number of (remaining) blocks at time t and D_t counts the cumulate number of frozen blocks until time t. This process jumps from (n, m) to $(n - k, m)$ when a coalescent event occurs (for some $k \in [n-1]$) and to $(n-1, m+1)$ when a freezing event occurs. In the special case of a Kingman coalescent with freezing, we use the notation $\{M_t^K, D_t\}_{t \geq 0}$.

We obtain the following analogue of Theorem 2.6.

Theorem 4.7 (Convergence of the window process I: Kingman limit). *Consider an extended seed bank di-graph with parameters $N, \mu^N, \mathcal{W}^N$ and $u_{N,1}, u_{N,2}$. Assume that conditions of Theorem 2.6 hold, plus the following*

$$u_{N,0} \to u_0, \qquad \frac{u_{N,1}\beta_N}{c_N} \to u_1, \qquad \frac{u_{N,2}\beta_N}{c_N} \to u_2,$$

where $u_0 \in (0, 1]$ and $u_1, u_2 > 0$. Consider the window process with mutations starting at $\bar{B}_0^N = \bar{n}$ and $D_0^N = 0$ for all $N \in \mathbb{Z}_+$ big enough. Then,

$$\lim_{N \to \infty} \{|\bar{B}_{\lfloor t\beta_N^2/c_N u_{N,0}^2\rfloor}^N|, D_{\lfloor t\beta_N^2/c_N u_{N,0}^2\rfloor}^N\}_{t \geq 0} = \{M_t^K, D_t\}_{t \geq 0} \qquad (4.3)$$

in the finite dimension sense, where $\{M_t^K, D_t\}_{t \geq 0}$ is the block counting process of a Kingman coalescent with freezing parameter $(u_1 + u_2)/u_0^2$.

Before proving this result, we need to modify the coupling particle system introduced in section 2.5. Let $Y_g^N = (R_g^N, L_g^N)$ define a Markov chain with state space $(\mathbb{N} \times [N]) \cup \{\Delta_1, \Delta_2\}$ and transition probabilities, conditional on the weights $\bar{W}_g^N$,

$$\mathbb{P}\big(Y_g^N = (i, k)|\{\bar{W}_g^N\}_g, Y_{g-1}^N = (1, j)\big) = W_{(g_0-g+1-i,k)}^N \mu^N(i)u_{N,0}$$

for every $i \geq 1$, $k \in [N]$,

$$\mathbb{P}\big(Y_g^N = \Delta_i|\{\bar{W}_g^N\}_g, Y_{g-1}^N = (1, j)\big) = u_{N,i}$$

for $i \in \{1, 2\}$, and

$$\mathbb{P}\big(Y_g^N = (i, k)|\{\bar{W}_g^N\}_g, Y_{g-1}^N = (i+1, j)\big) = W_{(g_0-g,k)}^N$$

for every $i \geq 1$ and $k \in [N]$. We enunciate the coupling result without proof.

Proposition 4.8. *Set $n = \sum n_i$ to be the total size of the initial sample. For every $g \geq 0$, consider n (conditional on $\{\bar{W}_g^N\}_g$) independent realizations of Y_g^N, that we call $Y_g^{N,j} = (R_g^{N,j}, L_g^{N,j})$ for $1 \leq j \leq n$, and set*

$$\gamma^{N,j} = \inf\{g \geq 1 : Y_g^{N,j} \in \{\Delta_1, \Delta_2\}\}.$$

Recall $\{\sigma^{N,j}\}_{j=1}^n$ from Proposition 2.5. For all $i \geq 1$, set $\sum_{j=1}^n 1_{\{R_0^{N,j}=i\}} = B_0^{N,i}$. Then, the i-th component $B_g^{N,i}$ of the random vector $\bar{B}_g^N$ is equal in distribution to $\sum_{j=1}^n 1_{\{R_g^{N,j}=i\}}1_{\{\sigma^{N,j}\geq g\}}1_{\{\gamma^{N,j}\geq g\}}$, for all $g \geq 0$ and D_g^N is equal in distribution to $\sum_{k=1}^g \sum_{j=1}^n 1_{\{\gamma^{N,j}=k\}}1_{\{\sigma^{N,j}\geq k\}}$.

We are now ready to prove Theorem 4.7.

Proof of Theorem 4.7: As in the proof of Theorem 2.6, recall the invariant measure ν^N defined in (2.2), define $\underline{Y}_g^{N,j} = (\underline{R}_g^{N,j}, L_g^{N,j})$ where $\{\underline{R}_g^{N,j}\}_{g \geq 0}$ is a sequence of independent ν^N-distributed random variables, and let $\underline{\sigma}^{N,1} = \infty$ and for $2 \leq j \leq n$,

$$\underline{\sigma}^{N,j} = \inf\{g \geq 1 : \underline{Y}_g^{N,j} = \underline{Y}_g^{N,j'} = (1,k), \text{ for some } j' < j \text{ such that } \underline{\sigma}^{N,j'} > g, k \in [N]\}.$$

To couple the variables $(\gamma^{N,j})_{j=1}^n$, observe that each of them can be associated to a geometric r.v. of parameter $u_{N,1} + u_{N,2}$. More precisely, let $(G^{N,1}, \dots, G^{N,n})$ be a family of independent geometric r.v.s of parameter $u_{N,1} + u_{N,2}$. Then we define

$$\underline{\gamma}^{N,j} = 1 + \inf\Big\{g \geq 1 : \sum_{k=0}^g 1_{\{\underline{R}_k^{N,j}=1\}} = G^{N,j}\Big\}.$$

Hence, we define an artificial window process with mutations $\{\bar{Z}_g^N, I_g^N\}_{g \geq 0}$ where

$$\bar{Z}_g^N = (Z_g^{N,1}, \dots, Z_g^{N,m_N})$$

has coordinates given by

$$Z_g^{N,i} = \sum_{j=1}^n 1_{\{\underline{R}_g^{N,j}=i\}} 1_{\{\underline{\sigma}^{N,j} \geq g\}} 1_{\{\underline{\gamma}^{N,j} \geq g\}}$$

and

$$I_g^N = \sum_{k=1}^g \sum_{j=1}^n 1_{\{\underline{\gamma}^{N,j}=k\}} 1_{\{\underline{\sigma}^{N,j} \geq k\}}.$$

The process $\{|\bar{Z}_g^N|, I_g^N\}_{g \geq 0}$ is Markovian.

First, we calculate the generator of $\{(|\bar{Z}_g^N|, I_g^N)\}_{g \geq 0}$ in order to discover its scaling limit. Let $f : \mathbb{Z}_+ \times \mathbb{Z}_+ \to \mathbb{R}^2$ be a bounded function. Then

$$
\begin{aligned}
\mathcal{G}^N f(n,m) &= \mathbb{E}[f(|\bar{Z}_1^N|, I_1^N) - f(n,m)] \\
&= \mathbb{P}((|\bar{Z}_1^N|, I_1^N) = (n-1,m))[f(n-1,m) - f(n,m)] \\
&\quad + \mathbb{P}((|\bar{Z}_1^N|, I_1^N) = (n-1,m+1))[f(n-1,m+1) - f(n,m)] \\
&\quad + O(\mathbb{P}(|\bar{Z}_1^N| = n-2)) \\
&= \binom{n}{2} c_N (\nu^N(1) u_{N,0})^2 [f(n-1,m) - f(n,m)] \\
&\quad + n\nu^N(1)(u_{N,1} + u_{N,2})[f(n-1,m+1) - f(n,m)] \\
&\quad + O\Big(d_N(\nu^N(1)u_{N,0})^3 + (\nu^N(1)(u_{N,1}+u_{N,2}))^2 + c_N(\nu^N(1))^3 u_{N,0}^2(u_{N,1}+u_{N,2})\Big) \\
&= \frac{c_N u_{N,0}^2}{\beta_N^2}\Big[\binom{n}{2}[f(n-1,m) - f(n,m)] \\
&\quad + n\left(\frac{u_{N,1}\beta_N}{c_N u_{N,0}^2} + \frac{u_{N,2}\beta_N}{c_N u_{N,0}^2}\right)[f(n-1,m+1) - f(n,m)] \\
&\quad + O\left(\frac{d_N}{\beta_N c_N}u_{N,0} + \frac{c_N}{\beta_N^2}\left(\frac{u_{N,1}\beta_N}{c_N u_{N,0}} + \frac{u_{N,2}\beta_N}{c_N u_{N,0}}\right)^2 + \frac{c_N}{\beta_N^2}\left(\frac{u_{N,1}\beta_N}{c_N} + \frac{u_{N,2}\beta_N}{c_N}\right)\right)\Big].
\end{aligned}
$$

So we conclude that

$$\{|\bar{Z}_{\lfloor t\beta_N^2/c_N u_{N,0}^2\rfloor}^N|, I_{\lfloor t\beta_N^2/c_N u_{N,0}^2\rfloor}^N\}_{t \geq 0} \Rightarrow \{M_t^K, D_t\}_{t \geq 0}, \tag{4.4}$$

the block counting process of a Kingman coalescent with freezing parameter $(u_1 + u_2)/u_0^2$.

To see that $\{|\bar{B}^N_{\lfloor t\beta_N^2/c_Nu_{N,o}^2\rfloor}|, D^N_{\lfloor t\beta_N^2/c_Nu_{N,0}^2\rfloor}\}_{t\geq0} \Rightarrow \{M_t^K, D_t\}_{t\geq0}$, we couple $\{|\bar{Z}^N_{\lfloor t\beta_N^2/c_N\rfloor}|\}_{t\geq0}$ and $\{|\bar{B}^N_{\lfloor t\beta_N^2/c_N\rfloor}|\}_{t\geq0}$ mimicking the proof of Theorem 2.6 to show that the same limit is true for the rescaled window process (since $u_{N,0} \to u_0$ with no scaling, we can suppose that $u_{N,0} = 1$). In this case, we still denote, for any $p, q \in [n]$, $\rho_k^{N,p,q} = \inf\{g > \rho_{k-1}^{N,p,q} : L_g^{N,p} = L_g^{N,q}\}$ (with $\rho_0^{N,p,q} = 0$), but now the potential jump times are $\rho_i = \inf\{g > \rho_{i-1} : g = \rho_k^{N,p,q}$ for some $p, q \in [n]$ and some $k \in \mathbb{Z}_+$ or $g = G^{N,j}$ for some $j \in [n]\}$ (with $\rho_0 = 0$). Recall that $(G^{N,1}, \ldots, G^{N,n})$ is a family of independent geometric r.v.s of parameter $u_{N,1} + u_{N,2}$. The probability that the coupling is successful is

$$p_N := \inf_{\bar{n}\in[N]^{m_N}} \mathbb{P}_{\bar{n}}((\underline{R}^{N,1}_{\hat{\rho}_1}, \ldots, \underline{R}^{N,m_N}_{\hat{\rho}_1}) = (R^{N,1}_{\hat{\rho}_1}, \ldots, R^{N,m_N}_{\hat{\rho}_1}))$$

$$= 1 - \sup_{\bar{n}\in[N]^{m_N}} \mathbb{P}_{\bar{n}}((\underline{R}^{N,1}_{\hat{\rho}_1}, \ldots, \underline{R}^{N,m_N}_{\hat{\rho}_1}) \neq (R^{N,1}_{\hat{\rho}_1}, \ldots, R^{N,m_N}_{\hat{\rho}_1}))$$

$$= 1 - \sup_{\bar{n}\in[N]^{m_N}} \|\mathbb{P}_{\bar{n}}((R^{N,1}_{\hat{\rho}_1}, \ldots, R^{N,m_N}_{\hat{\rho}_1}) = \cdot) - (\nu^N)^{\otimes m_N}(\cdot)\|_{TV}$$

where $\mathbb{P}_{\bar{n}}$ stands for the law of $\{R_g^{N,1}, \ldots, R_g^{N,m_N}\}_{g\geq0}$ (or $\{\underline{R}_g^{N,1}, \ldots, \underline{R}_g^{N,m_N}\}_{g\geq0}$) starting at the state $\bar{n} \in [N]^{m_N}$ and where Proposition 4.7 in Levin and Peres (2017) is used for the last equality. To prove that $p_N \to 1$ when $N \to \infty$, take $\varepsilon > 0$ such that $N^\varepsilon\tau_Nc_N \to 0$ (and thus $\tau_N N^\varepsilon(u_{N,1}+u_{N,2}) \to 0$) . The condition $\mu^N(1) > 0$ implies that, for any $i \geq 1$, the processes $\{R_g^{N,i}\}_{g\geq0}$ are irreducible. So, by Theorem 4.9 in Levin and Peres (2017), we have

$$\|\mathbb{P}_{\bar{n}}((R^{N,1}_{N^\varepsilon\tau_N}, \ldots, R^{N,m_N}_{N^\varepsilon\tau_N}) = \cdot) - (\nu^N)^{\otimes m_N}(\cdot)\|_{TV} < (1/4)^{N^\varepsilon}.$$

Then observe that, stochastically, $\rho_1 \geq \Gamma^N$ where Γ^N is a geometric random variable of parameter $n^2c_N + n(u_{N,1} + u_{N,2})$ and thus $\mathbb{P}(\rho_1 \leq N^\varepsilon\tau_N) \leq \mathbb{P}(\Gamma^N \leq N^\varepsilon\tau_N) \to 0$.

Let $T_1^N = \inf\{i \geq 1 : |\bar{Z}^N_{\rho_i}| = 1\}$. The jump times are reduced compared to those of the proof of Theorem 2.6, so the inequality (2.11) still holds and we obtain (4.3) with similar arguments ☐

A multiple merger version of this result is also obtained. Note that in this case the hypothesis that $\beta_N \to \beta < \infty$ involves that $u_{N,0} \to 1$.

Theorem 4.9 (Convergence of the window process II: Ξ limit). *Fix $\{\mu^N\}_{N\in\mathbb{N}}$ such that $\beta_N = \mathbb{E}[J_v^N] \to \beta < \infty$ and fix the distribution $\mathcal{W}^N$. Assume that the ancestral process of a Cannings model driven by $\mathcal{W}^N$, that we denote by $\{C_g^N\}_{g\geq0}$ is such that*

$$\lim_{N\to\infty} \{C^N_{\lfloor t/c_N\rfloor}\}_{t\geq0} = \{N_t^\Xi\}_{t\geq0}$$

where $\{N_t^\Xi\}_{t\geq0}$ stands for the block counting process of a Ξ-coalescent. Assume that

$$\frac{u_{N,1}}{\beta_Nc_N} \to u_1, \quad \frac{u_{N,2}}{\beta_Nc_N} \to u_2,$$

where $u_1, u_2 > 0$. Consider the window process with mutations starting at $\bar{B}_0^N = \bar{n}$ and $D_0^N = 0$ for all $N \in \mathbb{Z}_+$ big enough. Then,

$$\lim_{N\to\infty} \{|\bar{B}^N_{\lfloor t/c_N\rfloor}|, D^N_{\lfloor t/c_N\rfloor}\}_{t\geq0} = \{M_t^{\Xi^\beta}, D_t\}_{t\geq0}. \tag{4.5}$$

in the finite dimension sense, where $\{M_t^{\Xi^\beta}, D_t\}_{t\geq0}$ is the block counting process of a Ξ^β-coalescent with freezing parameter $u_1 + u_2$.

Finally we obtain a convergence result for the frequency processes. Because of the two coordinate notation of the block counting process of a coalescent with freezing, its moment dual is now written as $\{X_t, \theta_t\}_{t\geq0}$

Theorem 4.10 (Convergence of the forward frequency process)**.** *Assume that* $m_N \leq m < \infty$ *for all* $N \in \mathbb{Z}_+$. *Fix* $\{\mathcal{W}^N\}_{N \geq 1}$, $\{\mu^N\}_{N \geq 1}$ *(and the associated stationary distribution* ν^N*) and* $u_{N,1}, u_{N,2}$ *such that either the assumptions of Theorem 4.7 hold or the assumptions of Theorem 4.9 hold. Suppose that* ν^N *converges to a measure* ν *on* $[m]$ *as* $N \to \infty$. *Let* $\{\bar{X}^N, \theta^N\}_{N \geq 1}$ *be the sequence of frequency processes with parameters* N, $\mathcal{W}^N$, μ^N *and* $u_{N,1}, u_{N,2}$ *and starting condition* $\bar{X}_0^N = (\lfloor Nx_1 \rfloor/N, \ldots, \lfloor Nx_m \rfloor/N)$ *for some* $\bar{x} \in [0,1]^m$. *i) Under that assumptions of Theorem 4.7,*

$$\lim_{N \to \infty} \{\bar{X}^N_{\lfloor t\beta_N^2/c_N \rfloor}\}_{t \geq 0} = \{\bar{X}_t\}_{t \geq 0}$$

in the finite dimension sense, where $\bar{X}_t$ *is a vector with* m *identical coordinates* X_t *such that* $X_0 = x_0 = \sum_{i=1}^m \nu(i)x_i$ *a.s., and* $\{X_t, u_1/(u_1 + u_2)\}_{t \geq 0}$ *is the moment dual of* $\{M_t^K, D_t\}_{t \geq 0}$.
ii) Under the assumptions of Theorem 4.9,

$$\lim_{N \to \infty} \{\bar{X}^N_{\lfloor t/c_N \rfloor}\}_{t \geq 0} = \{\bar{X}_t\}_{t \geq 0}$$

in the finite dimension sense, where $\bar{X}_t$ *is a vector with* m *identical coordinates* X_t *such that* $X_0 = x_0 = \sum_{i=1}^m \nu(i)x_i$ *a.s., and* $\{X_t, u_1/(u_1 + u_2)\}_{t \geq 0}$ *is moment dual of* $\{M_t^{\Xi^\beta}, D_t\}_{t \geq 0}$.

Acknowledgements

References

Birkner, M., Liu, H., and Sturm, A. Coalescent results for diploid exchangeable population models. *Electron. J. Probab.*, **23**, Paper No. 49, 44 (2018). MR3814243.

Blath, J., Buzzoni, E., González Casanova, A., and Wilke-Berenguer, M. Structural properties of the seed bank and the two island diffusion. *J. Math. Biol.*, **79** (1), 369–392 (2019). MR3975877.

Blath, J., Eldon, B., González Casanova, A., and Kurt, N. Genealogy of a Wright-Fisher model with strong seedbank component. In *XI Symposium on Probability and Stochastic Processes*, volume 69 of *Progr. Probab.*, pp. 81–100. Birkhäuser/Springer, Cham (2015). MR3558137.

Blath, J., González Casanova, A., Kurt, N., and Spanò, D. The ancestral process of long-range seed bank models. *J. Appl. Probab.*, **50** (3), 741–759 (2013). MR3102512.

Blath, J., González Casanova, A., Kurt, N., and Wilke-Berenguer, M. A new coalescent for seed-bank models. *Ann. Appl. Probab.*, **26** (2), 857–891 (2016). MR3476627.

Blath, J., González Casanova, A., Kurt, N., and Wilke-Berenguer, M. The seed bank coalescent with simultaneous switching. *Electron. J. Probab.*, **25**, Paper No. 27, 21 (2020). MR4073688.

Blath, J., Hammer, M., and Nie, F. The stochastic Fisher-KPP equation with seed bank and on/off branching coalescing Brownian motion. *Stoch. Partial Differ. Equ. Anal. Comput.*, **11** (2), 773–818 (2023). MR4588623.

Cannings, C. The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. *Advances in Appl. Probability*, **6**, 260–290 (1974). MR343949.

Cannings, C. The latent roots of certain Markov chains arising in genetics: a new approach. II. Further haploid models. *Advances in Appl. Probability*, **7**, 264–282 (1975). MR371430.

Etheridge, A. and Griffiths, R. A coalescent dual process in a Moran model with genic selection. *Theoret. Population Biol.*, **75** (4), 320–330 (2009). Sam Karlin: Special Issue. DOI: 10.1016/j.tpb.2009.03.004.

Etheridge, A. M., Griffiths, R. C., and Taylor, J. E. A coalescent dual process in a Moran model with genic selection, and the lambda coalescent limit. *Theoret. Population Biol.*, **78** (2), 77–92 (2010). DOI: 10.1016/j.tpb.2010.05.004.

Ethier, S. N. and Griffiths, R. C. The transition function of a Fleming-Viot process. *Ann. Probab.*, **21** (3), 1571–1590 (1993). MR1235429.

Ewens, W. J. The sampling theory of selectively neutral alleles. *Theoret. Population Biol.*, **3** (1), 87–112 (1972). MR325177.

Freund, F. Cannings models, population size changes and multiple-merger coalescents. *J. Math. Biol.*, **80** (5), 1497–1521 (2020). MR4071423.

González Casanova, A., Miró Pina, V., and Siri-Jégousse, A. The symmetric coalescent and Wright-Fisher models with bottlenecks. *Ann. Appl. Probab.*, **32** (1), 235–268 (2022a). MR4386526.

González Casanova, A., Peñaloza, L., and Siri-Jégousse, A. The shape of a seed bank tree. *J. Appl. Probab.*, **59** (3), 631–651 (2022b). MR4480072.

González Casanova, A. and Spanò, D. Duality and fixation in Ξ-Wright-Fisher processes with frequency-dependent selection. *Ann. Appl. Probab.*, **28** (1), 250–284 (2018). MR3770877.

Griffiths, R. C. and Spanó, D. Diffusion processes and coalescent trees. In *Probability and mathematical genetics*, volume 378 of *London Math. Soc. Lecture Note Ser.*, pp. 358–379. Cambridge Univ. Press, Cambridge (2010). MR2744247.

Hammond, A. and Sheffield, S. Power law Pólya's urn and fractional Brownian motion. *Probab. Theory Related Fields*, **157** (3-4), 691–719 (2013). MR3129801.

Hobolth, A., Siri-Jégousse, A., and Bladt, M. Phase-type distributions in population genetics. *Theoret. Population Biol.*, **127**, 16–32 (2019). DOI: 10.1016/j.tpb.2019.02.001.

Igelbrink, J. L. and Wakolbinger, A. Asymptotic Gaussianity via coalescence probabilities in the Hammond-Sheffield urn. *ALEA Lat. Am. J. Probab. Math. Stat.*, **20** (1), 53–74 (2023). MR4535240.

Kaj, I., Krone, S. M., and Lascoux, M. Coalescent theory for seed bank models. *J. Appl. Probab.*, **38** (2), 285–300 (2001). MR1834743.

Kallenberg, O. *Foundations of modern probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, Cham, third edition (2021). ISBN 978-3-030-61871-1; 978-3-030-61870-4. MR4226142.

Kersting, G., Siri-Jégousse, A., and Wences, A. H. Site frequency spectrum of the Bolthausen-Sznitman coalescent. *ALEA Lat. Am. J. Probab. Math. Stat.*, **18** (2), 1483–1505 (2021). MR4282195.

Koopmann, B., Müller, J., Tellier, A., and Živković, D. Fisher–Wright model with deterministic seed bank and selection. *Theoret. Population Biol.*, **114**, 29–39 (2017). DOI: 10.1016/j.tpb.2016.11.005.

Lambert, A. and Ma, C. The coalescent in peripatric metapopulations. *J. Appl. Probab.*, **52** (2), 538–557 (2015). MR3372091.

Levin, D. A. and Peres, Y. *Markov chains and mixing times.* American Mathematical Society, Providence, RI, second edition (2017). ISBN 978-1-4704-2962-1. MR3726904.

Möhle, M. and Sagitov, S. A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.*, **29** (4), 1547–1562 (2001). MR1880231.

Sagitov, S. The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.*, **36** (4), 1116–1125 (1999). MR1742154.

Schweinsberg, J. Coalescent processes obtained from supercritical Galton-Watson processes. *Stochastic Process. Appl.*, **106** (1), 107–139 (2003). MR1983046.

Schweinsberg, J. R. *Coalescents with simultaneous multiple collisions.* ProQuest LLC, Ann Arbor, MI (2001). ISBN 978-0493-31022-0. PhD Thesis – University of California, Berkeley. MR2702298.

Siri-Jégousse, A. and Wences, A. H. Exchangeable coalescents beyond the Cannings class. *ArXiv Mathematics e-prints* (2022). arXiv: 2212.02154.