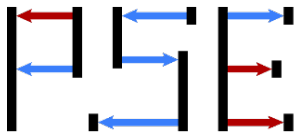
Genetic variability under the seed bank coalescent

Jochen Blath

Joint work with

Bjarki Eldon, Adrián González Casanova, Noemi Kurt, Maite Wilke Berenguer
(all TU Berlin)

CIRM Luminy, June 2015



PROBABILISTIC STRUCTURES
IN EVOLUTION

DFG SPP 1590

Seed banks in population genetics

A variety of species produce *seeds* or *dormant forms* which introduce strong age-structure / *seed banks* in population genetic models.

- Seed banks can act as *buffer* against evolutionary forces such as random genetic drift and selection; 'bet-hedging' strategy to overcome unfavourable environmental conditions.
- Their presence typically leads to significantly *increased genetic variability*.
- Classical mechanisms such as *fixation* and *extinction* become *more complex*: Genetic types can disappear from the active population while returning later due to the germination of seeds or activation of dormant forms.

Seed banks in population genetics

A variety of species produce *seeds* or *dormant forms* which introduce strong age-structure / *seed banks* in population genetic models.

- Seed banks can act as *buffer* against evolutionary forces such as random genetic drift and selection; 'bet-hedging' strategy to overcome unfavourable environmental conditions.
- Their presence typically leads to significantly *increased genetic variability*.
- Classical mechanisms such as *fixation* and *extinction* become *more complex*: Genetic types can disappear from the active population while returning later due to the germination of seeds or activation of dormant forms.

Seed banks in population genetics

A variety of species produce *seeds* or *dormant forms* which introduce strong age-structure / *seed banks* in population genetic models.

- Seed banks can act as *buffer* against evolutionary forces such as random genetic drift and selection; 'bet-hedging' strategy to overcome unfavourable environmental conditions.
- Their presence typically leads to significantly *increased genetic variability*.
- Classical mechanisms such as *fixation* and *extinction* become *more complex*: Genetic types can disappear from the active population while returning later due to the germination of seeds or activation of dormant forms.

Seed banks in population genetics

A variety of species produce *seeds* or *dormant forms* which introduce strong age-structure / *seed banks* in population genetic models.

- Seed banks can act as *buffer* against evolutionary forces such as random genetic drift and selection; 'bet-hedging' strategy to overcome unfavourable environmental conditions.
- Their presence typically leads to significantly *increased genetic variability*.
- Classical mechanisms such as *fixation* and *extinction* become *more complex*: Genetic types can disappear from the active population while returning later due to the germination of seeds or activation of dormant forms.

Dormancy in microbial populations

Seed bank effects have also been suggested to play an important role in microbial evolution [Lennon & Jones, *Nature reviews*, 2011]:

- Many microbial species exhibit *dormant forms*. That is, organisms can enter (and leave) a reversible state of low (resp. vanishing) metabolic activity. These forms can be short-lived but may also stay inactive for significant periods of time. A variety of bacteria can produce endospores or cysts that remain viable for many decades/centuries.
- Dormant microorganisms generate a *seed bank*, which comprises inactive individuals that are capable of being resuscitated.
- Initiation of dormancy may be triggered by environment, but may also happen spontaneously (*responsive* vs. *spontaneous switching*).
- A large fraction of the microorganisms in nature seem to be metabolically inactive.

Dormancy in microbial populations

Seed bank effects have also been suggested to play an important role in microbial evolution [Lennon & Jones, *Nature reviews*, 2011]:

- Many microbial species exhibit *dormant forms*. That is, organisms can enter (and leave) a reversible state of low (resp. vanishing) metabolic activity. These forms can be short-lived but may also stay inactive for significant periods of time. A variety of bacteria can produce endospores or cysts that remain viable for many decades/centuries.
- Dormant microorganisms generate a *seed bank*, which comprises inactive individuals that are capable of being resuscitated.
- Initiation of dormancy may be triggered by environment, but may also happen spontaneously (*responsive* vs. *spontaneous switching*).
- A large fraction of the microorganisms in nature seem to be metabolically inactive.

Dormancy in microbial populations

Seed bank effects have also been suggested to play an important role in microbial evolution [Lennon & Jones, *Nature reviews*, 2011]:

- Many microbial species exhibit *dormant forms*. That is, organisms can enter (and leave) a reversible state of low (resp. vanishing) metabolic activity. These forms can be short-lived but may also stay inactive for significant periods of time. A variety of bacteria can produce endospores or cysts that remain viable for many decades/centuries.
- Dormant microorganisms generate a *seed bank*, which comprises inactive individuals that are capable of being resuscitated.
- Initiation of dormancy may be triggered by environment, but may also happen spontaneously (*responsive* vs. *spontaneous switching*).
- A large fraction of the microorganisms in nature seem to be metabolically inactive.

Dormancy in microbial populations

Seed bank effects have also been suggested to play an important role in microbial evolution [Lennon & Jones, *Nature reviews*, 2011]:

- Many microbial species exhibit *dormant forms*. That is, organisms can enter (and leave) a reversible state of low (resp. vanishing) metabolic activity. These forms can be short-lived but may also stay inactive for significant periods of time. A variety of bacteria can produce endospores or cysts that remain viable for many decades/centuries.
- Dormant microorganisms generate a *seed bank*, which comprises inactive individuals that are capable of being resuscitated.
- Initiation of dormancy may be triggered by environment, but may also happen spontaneously (*responsive* vs. *spontaneous switching*).
- A large fraction of the microorganisms in nature seem to be metabolically inactive.

Dormancy in microbial communities

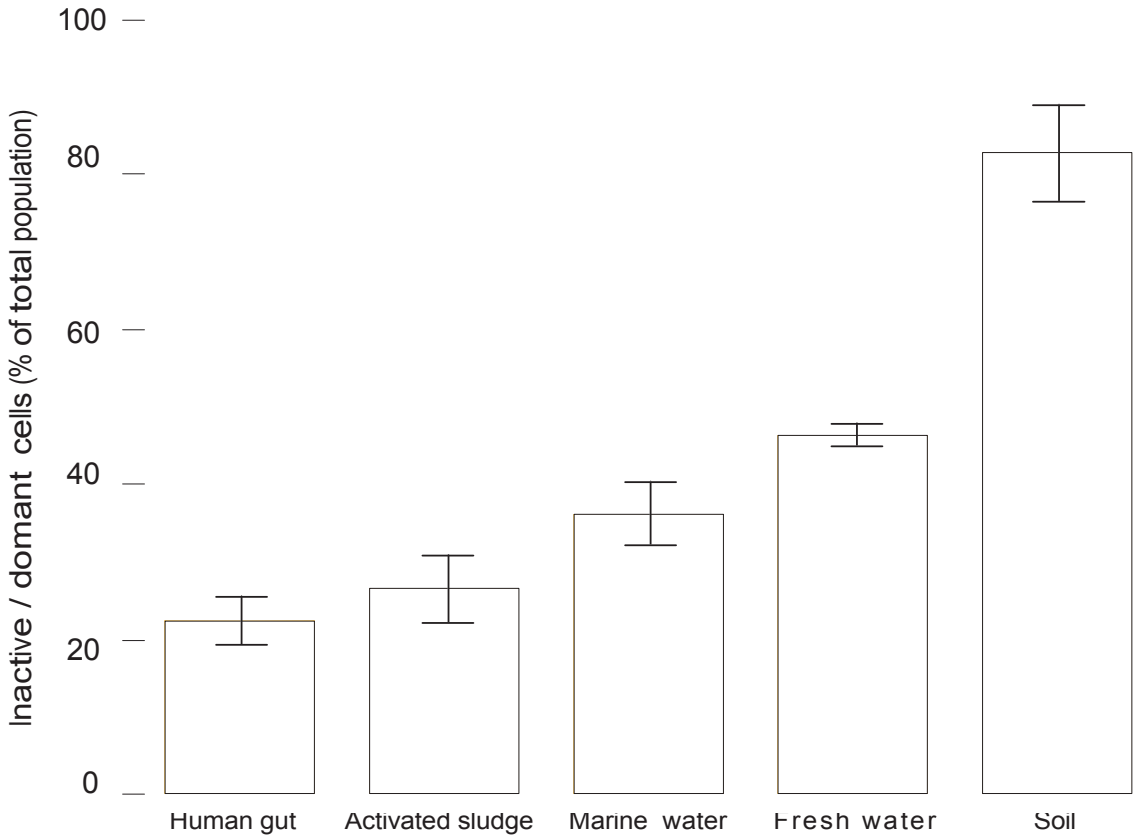


Figure: Percentage of inactive cells in microbial communities, data from [Lennon & Jones, 2011])

Modeling and investigating seed-banks

Despite many empirical studies and several theoretical works (including e.g. [Kaj, Krone & Lascoux 2001], [Vitalis, Glémin, Oliveiri 2004], [Tellier, Laurent, Lainer, Pavlidis, Stephan 2011], [Živković & Tellier 2011],...), the mathematical modeling of seed banks in population genetics appears to be still incomplete.

Aim of this talk: Include large seed banks with potentially extended periods of dormancy in classical *Wright-Fisher population models*; obtain scaling limits; investigate ancestral relationships in terms of coalescent processes, derive expressions for population genetic quantities to describe genetic variability.

Modeling and investigating seed-banks

Despite many empirical studies and several theoretical works (including e.g. [Kaj, Krone & Lascoux 2001], [Vitalis, Glémin, Oliveiri 2004], [Tellier, Laurent, Lainer, Pavlidis, Stephan 2011], [Živković & Tellier 2011],...), the mathematical modeling of seed banks in population genetics appears to be still incomplete.

Aim of this talk: Include large seed banks with potentially extended periods of dormancy in classical *Wright-Fisher population models*; obtain scaling limits; investigate ancestral relationships in terms of coalescent processes, derive expressions for population genetic quantities to describe genetic variability.

Known results, I

- [Kaj, Krone & Lascoux 2001] include seed bank effect in a classical Wright-Fisher model: In a population of size N, each individual independently picks its parent uniformly from a randomly chosen previous generation B (with law μ) back in time.
- They show that if μ is supported on $\{1, 2, \ldots, m\}$ (*independent* of N), then the ancestral process converges, as $N \to \infty$, after the usual time-scaling by N, to a (time-changed) *Kingman coalescent*, with coalescence rates multiplied by the constant $\beta^2 := 1/\mathbb{E}[B]^2$.
- An increase of $E[B]$ thus *decelerates* the coalescent, leading to an increase in the effective population size.
- However, in a set-up with neutral mutation, since the overall coalescent tree structure is retained, this leaves the relative genetic type frequencies in the normalized site frequency spectrum unchanged. In this case, we speak of a *'weak' seed bank effect*.

Known results, I

- [Kaj, Krone & Lascoux 2001] include seed bank effect in a classical Wright-Fisher model: In a population of size N, each individual independently picks its parent uniformly from a randomly chosen previous generation B (with law μ) back in time.
- They show that if μ is supported on $\{1, 2, \ldots, m\}$ (*independent* of N), then the ancestral process converges, as $N \to \infty$, after the usual time-scaling by N, to a (time-changed) *Kingman coalescent*, with coalescence rates multiplied by the constant $\beta^2 := 1/\mathbb{E}[B]^2$.
- An increase of $E[B]$ thus *decelerates* the coalescent, leading to an increase in the effective population size.
- However, in a set-up with neutral mutation, since the overall coalescent tree structure is retained, this leaves the relative genetic type frequencies in the normalized site frequency spectrum unchanged. In this case, we speak of a *'weak' seed bank effect*.

Known results, I

- [Kaj, Krone & Lascoux 2001] include seed bank effect in a classical Wright-Fisher model: In a population of size N, each individual independently picks its parent uniformly from a randomly chosen previous generation B (with law μ) back in time.
- They show that if μ is supported on $\{1, 2, \ldots, m\}$ (*independent* of N), then the ancestral process converges, as $N \to \infty$, after the usual time-scaling by N, to a (time-changed) *Kingman coalescent*, with coalescence rates multiplied by the constant $\beta^2 := 1/\mathbb{E}[B]^2$.
- An increase of $E[B]$ thus *decelerates* the coalescent, leading to an increase in the effective population size.
- However, in a set-up with neutral mutation, since the overall coalescent tree structure is retained, this leaves the relative genetic type frequencies in the normalized site frequency spectrum unchanged. In this case, we speak of a *'weak' seed bank effect*.

Known results, I

- [Kaj, Krone & Lascoux 2001] include seed bank effect in a classical Wright-Fisher model: In a population of size N, each individual independently picks its parent uniformly from a randomly chosen previous generation B (with law μ) back in time.
- They show that if μ is supported on $\{1, 2, \ldots, m\}$ (*independent* of N), then the ancestral process converges, as $N \to \infty$, after the usual time-scaling by N, to a (time-changed) *Kingman coalescent*, with coalescence rates multiplied by the constant $\beta^2 := 1/\mathbb{E}[B]^2$.
- An increase of $E[B]$ thus *decelerates* the coalescent, leading to an increase in the effective population size.
- However, in a set-up with neutral mutation, since the overall coalescent tree structure is retained, this leaves the relative genetic type frequencies in the normalized site frequency spectrum unchanged. In this case, we speak of a *'weak' seed bank effect*.

Known results, II

- More generally, [B., González, Kurt, Spanò 2013] show that a sufficient condition for convergence to the Kingman coalescent (with similar scaling and rates) is that $E[B] < \infty$ (again, with B resp. μ independent of N).
- Further, they show that an extreme seed-bank effect can completely alter genealogical behaviour: If the seed bank age distribution μ is *heavy-tailed*, say,

$$\mu(k) = L(k)k^{-\alpha},$$

  where L is slowly varying, then, if $\alpha < 1$, the expected time to the most recent common ancestor is infinite, and if $\alpha < 1/2$ two randomly sampled individuals do not have a common ancestor at all wpp.
- Some related models have been investigated in [B., Eldon, González, Kurt 2015], but they all either lead to a Kingman coalescent (potentially on different time-scales) or degenerate ancestral processes.
- Problem: Long-range seed-bank models are *highly non-Markovian!*

Known results, II

- More generally, [B., González, Kurt, Spanò 2013] show that a sufficient condition for convergence to the Kingman coalescent (with similar scaling and rates) is that $E[B] < \infty$ (again, with B resp. μ independent of N).
- Further, they show that an extreme seed-bank effect can completely alter genealogical behaviour: If the seed bank age distribution μ is *heavy-tailed*, say,

$$\mu(k) = L(k)k^{-\alpha},$$

  where L is slowly varying, then, if $\alpha < 1$, the expected time to the most recent common ancestor is infinite, and if $\alpha < 1/2$ two randomly sampled individuals do not have a common ancestor at all wpp.
- Some related models have been investigated in [B., Eldon, González, Kurt 2015], but they all either lead to a Kingman coalescent (potentially on different time-scales) or degenerate ancestral processes.
- Problem: Long-range seed-bank models are *highly non-Markovian!*

Known results, II

- More generally, [B., González, Kurt, Spanò 2013] show that a sufficient condition for convergence to the Kingman coalescent (with similar scaling and rates) is that $E[B] < \infty$ (again, with B resp. μ independent of N).
- Further, they show that an extreme seed-bank effect can completely alter genealogical behaviour: If the seed bank age distribution μ is *heavy-tailed*, say,

$$\mu(k) = L(k)k^{-\alpha},$$

  where L is slowly varying, then, if $\alpha < 1$, the expected time to the most recent common ancestor is infinite, and if $\alpha < 1/2$ two randomly sampled individuals do not have a common ancestor at all wpp.
- Some related models have been investigated in [B., Eldon, González, Kurt 2015], but they all either lead to a Kingman coalescent (potentially on different time-scales) or degenerate ancestral processes.
- Problem: Long-range seed-bank models are *highly non-Markovian!*

Known results, II

- More generally, [B., González, Kurt, Spanò 2013] show that a sufficient condition for convergence to the Kingman coalescent (with similar scaling and rates) is that $E[B] < \infty$ (again, with B resp. μ independent of N).
- Further, they show that an extreme seed-bank effect can completely alter genealogical behaviour: If the seed bank age distribution μ is *heavy-tailed*, say,

$$\mu(k) = L(k)k^{-\alpha},$$

  where L is slowly varying, then, if $\alpha < 1$, the expected time to the most recent common ancestor is infinite, and if $\alpha < 1/2$ two randomly sampled individuals do not have a common ancestor at all wpp.
- Some related models have been investigated in [B., Eldon, González, Kurt 2015], but they all either lead to a Kingman coalescent (potentially on different time-scales) or degenerate ancestral processes.
- Problem: Long-range seed-bank models are *highly non-Markovian!*

The Wright-Fisher model with large geometric seed bank component

In this talk, we investigate a simple Markovian seed bank model, where the seed bank size is comparable to the original population N, and where the average dormancy period may also be of order N. This leads to a natural new ancestral scaling limit, which we call *'seed bank coalescent'*.

The Wright-Fisher model with geometric seed bank component - notation

Consider a haploid population of fixed size N reproducing in fixed discrete generations $k = 0, 1, \ldots$. Assume that each individual carries a genetic type from some type-space E, say $E = \{a, A\}$.

Further, assume that the population also sustains a *seed bank* of constant size M in each generation, which consists of the dormant individuals. For simplicity, we 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 $\delta := \varepsilon N/M$ (i.e., $\delta M = \varepsilon N$), and assume for convenience that all involved products and fractions are integers. Let $c = \varepsilon N$.

The Wright-Fisher model with geometric seed bank component - notation

Consider a haploid population of fixed size N reproducing in fixed discrete generations $k = 0, 1, \ldots$. Assume that each individual carries a genetic type from some type-space E, say $E = \{a, A\}$.

Further, assume that the population also sustains a *seed bank* of constant size M in each generation, which consists of the dormant individuals. For simplicity, we 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 $\delta := \varepsilon N/M$ (i.e., $\delta M = \varepsilon N$), and assume for convenience that all involved products and fractions are integers. Let $c = \varepsilon N$.

The Wright-Fisher model with geometric seed bank component - notation

Consider a haploid population of fixed size N reproducing in fixed discrete generations $k = 0, 1, \ldots$. Assume that each individual carries a genetic type from some type-space E, say $E = \{a, A\}$.

Further, assume that the population also sustains a *seed bank* of constant size M in each generation, which consists of the dormant individuals. For simplicity, we 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 $\delta := \varepsilon N/M$ (i.e., $\delta M = \varepsilon N$), and assume for convenience that all involved products and fractions are integers. Let $c = \varepsilon N$.

The Wright-Fisher model with geometric seed bank component - dynamics

- The N *plants* from generation 0 produce $(1-\varepsilon)N$ *plants* in generation 1 by multinomial sampling with equal weights (ordinary WF dynamics).
- Additionally, $\delta M = \varepsilon N$ uniformly sampled *seeds* from the seed-bank of size M in generation 0 'germinate', that is, they turn into exactly one *plant* in generation 1 each, and thus vacate the seed-bank.
- The *plants* from generation 0 are thus replaced by these $(1-\varepsilon)N + \delta M = N$ new active individuals, forming the *plants* in generation 1.
- For the seed-bank, the N *plants* from generation 0 produce $\delta M = \varepsilon N$ *seeds* by multinomial sampling, replacing those *seeds* that germinated.
- The remaining $(1-\delta)M$ *seeds* from generation 0 remain inactive and stay in the seed-bank.
- Throughout reproduction, offspring and seeds copy/resp. maintain the genetic type of the parent.

Thus, in generation 1, we have again N *plants* and M *seeds*. This probabilistic mechanism is then to be repeated independently in generations $k = 2, 3, ...$

The Wright-Fisher model with geometric seed bank component - dynamics

- The N *plants* from generation 0 produce $(1-\varepsilon)N$ *plants* in generation 1 by multinomial sampling with equal weights (ordinary WF dynamics).
- Additionally, $\delta M = \varepsilon N$ uniformly sampled *seeds* from the seed-bank of size M in generation 0 'germinate', that is, they turn into exactly one *plant* in generation 1 each, and thus vacate the seed-bank.
- The *plants* from generation 0 are thus replaced by these $(1-\varepsilon)N + \delta M = N$ new active individuals, forming the *plants* in generation 1.
- For the seed-bank, the N *plants* from generation 0 produce $\delta M = \varepsilon N$ *seeds* by multinomial sampling, replacing those *seeds* that germinated.
- The remaining $(1-\delta)M$ *seeds* from generation 0 remain inactive and stay in the seed-bank.
- Throughout reproduction, offspring and seeds copy/resp. maintain the genetic type of the parent.

Thus, in generation 1, we have again N *plants* and M *seeds*. This probabilistic mechanism is then to be repeated independently in generations $k = 2, 3, ...$

The Wright-Fisher model with geometric seed bank component - dynamics

- The N *plants* from generation 0 produce $(1-\varepsilon)N$ *plants* in generation 1 by multinomial sampling with equal weights (ordinary WF dynamics).
- Additionally, $\delta M = \varepsilon N$ uniformly sampled *seeds* from the seed-bank of size M in generation 0 'germinate', that is, they turn into exactly one *plant* in generation 1 each, and thus vacate the seed-bank.
- The *plants* from generation 0 are thus replaced by these $(1-\varepsilon)N + \delta M = N$ new active individuals, forming the *plants* in generation 1.
- For the seed-bank, the N *plants* from generation 0 produce $\delta M = \varepsilon N$ *seeds* by multinomial sampling, replacing those *seeds* that germinated.
- The remaining $(1-\delta)M$ *seeds* from generation 0 remain inactive and stay in the seed-bank.
- Throughout reproduction, offspring and seeds copy/resp. maintain the genetic type of the parent.

Thus, in generation 1, we have again N *plants* and M *seeds*. This probabilistic mechanism is then to be repeated independently in generations $k = 2, 3, ...$

The Wright-Fisher model with geometric seed bank component - dynamics

- The N *plants* from generation 0 produce $(1-\varepsilon)N$ *plants* in generation 1 by multinomial sampling with equal weights (ordinary WF dynamics).
- Additionally, $\delta M = \varepsilon N$ uniformly sampled *seeds* from the seed-bank of size M in generation 0 'germinate', that is, they turn into exactly one *plant* in generation 1 each, and thus vacate the seed-bank.
- The *plants* from generation 0 are thus replaced by these $(1-\varepsilon)N + \delta M = N$ new active individuals, forming the *plants* in generation 1.
- For the seed-bank, the N *plants* from generation 0 produce $\delta M = \varepsilon N$ *seeds* by multinomial sampling, replacing those *seeds* that germinated.
- The remaining $(1-\delta)M$ *seeds* from generation 0 remain inactive and stay in the seed-bank.
- Throughout reproduction, offspring and seeds copy/resp. maintain the genetic type of the parent.

Thus, in generation 1, we have again N *plants* and M *seeds*. This probabilistic mechanism is then to be repeated independently in generations $k = 2, 3, ...$

The Wright-Fisher model with geometric seed bank component - dynamics

- The N *plants* from generation 0 produce $(1-\varepsilon)N$ *plants* in generation 1 by multinomial sampling with equal weights (ordinary WF dynamics).
- Additionally, $\delta M = \varepsilon N$ uniformly sampled *seeds* from the seed-bank of size M in generation 0 'germinate', that is, they turn into exactly one *plant* in generation 1 each, and thus vacate the seed-bank.
- The *plants* from generation 0 are thus replaced by these $(1-\varepsilon)N + \delta M = N$ new active individuals, forming the *plants* in generation 1.
- For the seed-bank, the N *plants* from generation 0 produce $\delta M = \varepsilon N$ *seeds* by multinomial sampling, replacing those *seeds* that germinated.
- The remaining $(1-\delta)M$ *seeds* from generation 0 remain inactive and stay in the seed-bank.
- Throughout reproduction, offspring and seeds copy/resp. maintain the genetic type of the parent.

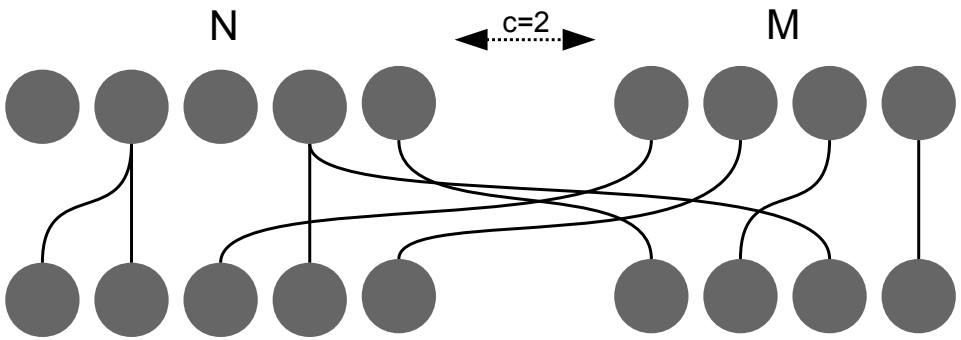
Thus, in generation 1, we have again N *plants* and M *seeds*. This probabilistic mechanism is then to be repeated independently in generations $k = 2, 3, ...$

The Wright-Fisher model with geometric seed bank component - dynamics

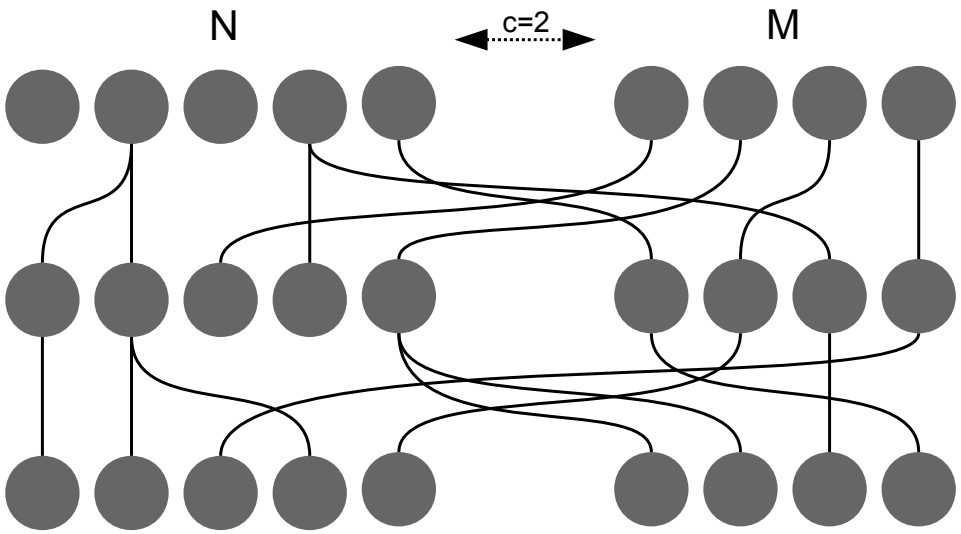
- The N *plants* from generation 0 produce $(1-\varepsilon)N$ *plants* in generation 1 by multinomial sampling with equal weights (ordinary WF dynamics).
- Additionally, $\delta M = \varepsilon N$ uniformly sampled *seeds* from the seed-bank of size M in generation 0 'germinate', that is, they turn into exactly one *plant* in generation 1 each, and thus vacate the seed-bank.
- The *plants* from generation 0 are thus replaced by these $(1-\varepsilon)N + \delta M = N$ new active individuals, forming the *plants* in generation 1.
- For the seed-bank, the N *plants* from generation 0 produce $\delta M = \varepsilon N$ *seeds* by multinomial sampling, replacing those *seeds* that germinated.
- The remaining $(1-\delta)M$ *seeds* from generation 0 remain inactive and stay in the seed-bank.
- Throughout reproduction, offspring and seeds copy/resp. maintain the genetic type of the parent.

Thus, in generation 1, we have again N *plants* and M *seeds*. This probabilistic mechanism is then to be repeated independently in generations $k = 2, 3, ...$

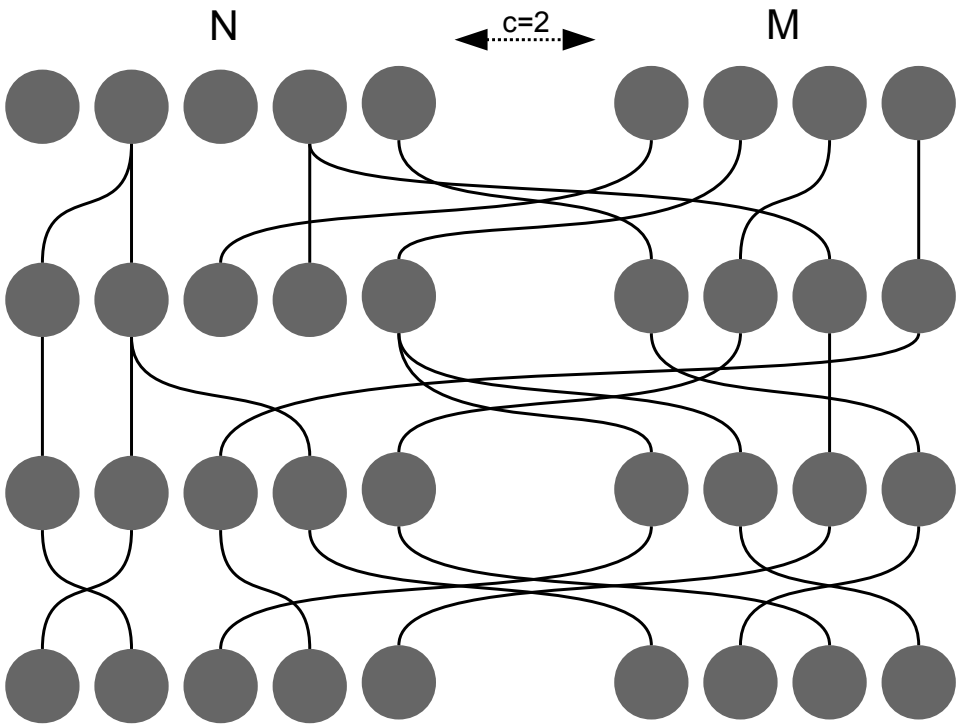
The Wright-Fisher model with geometric seed-bank



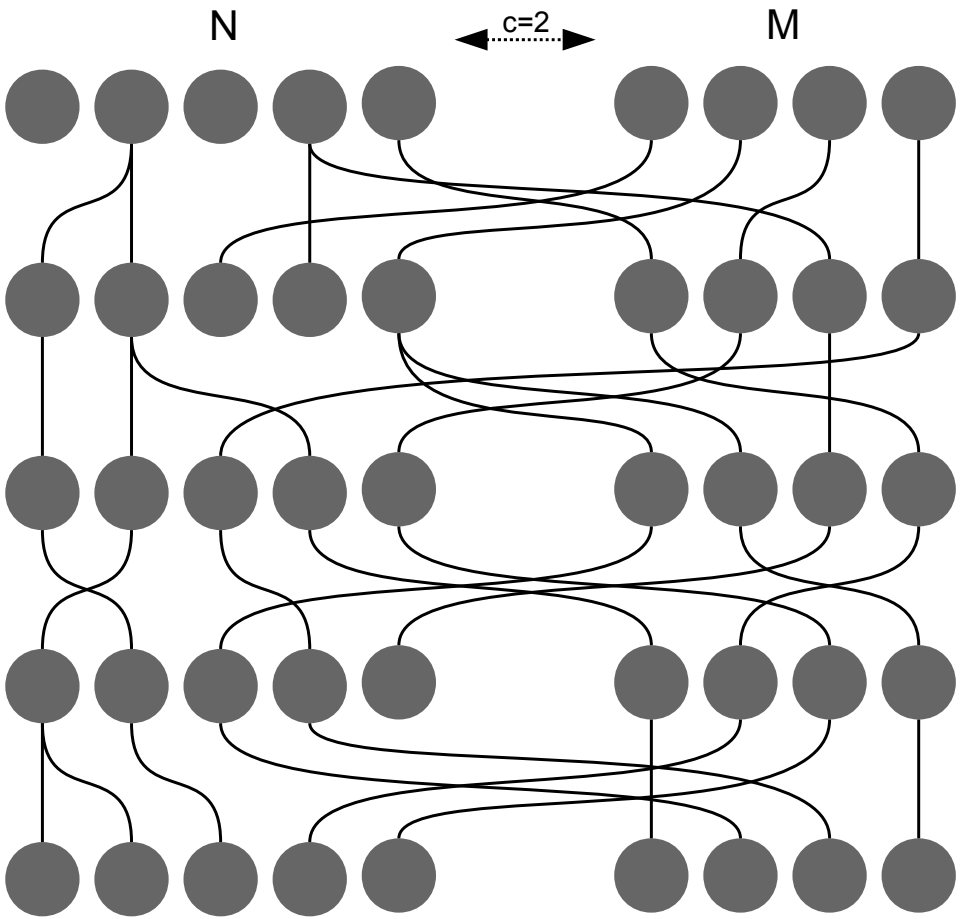
The Wright-Fisher model with geometric seed-bank



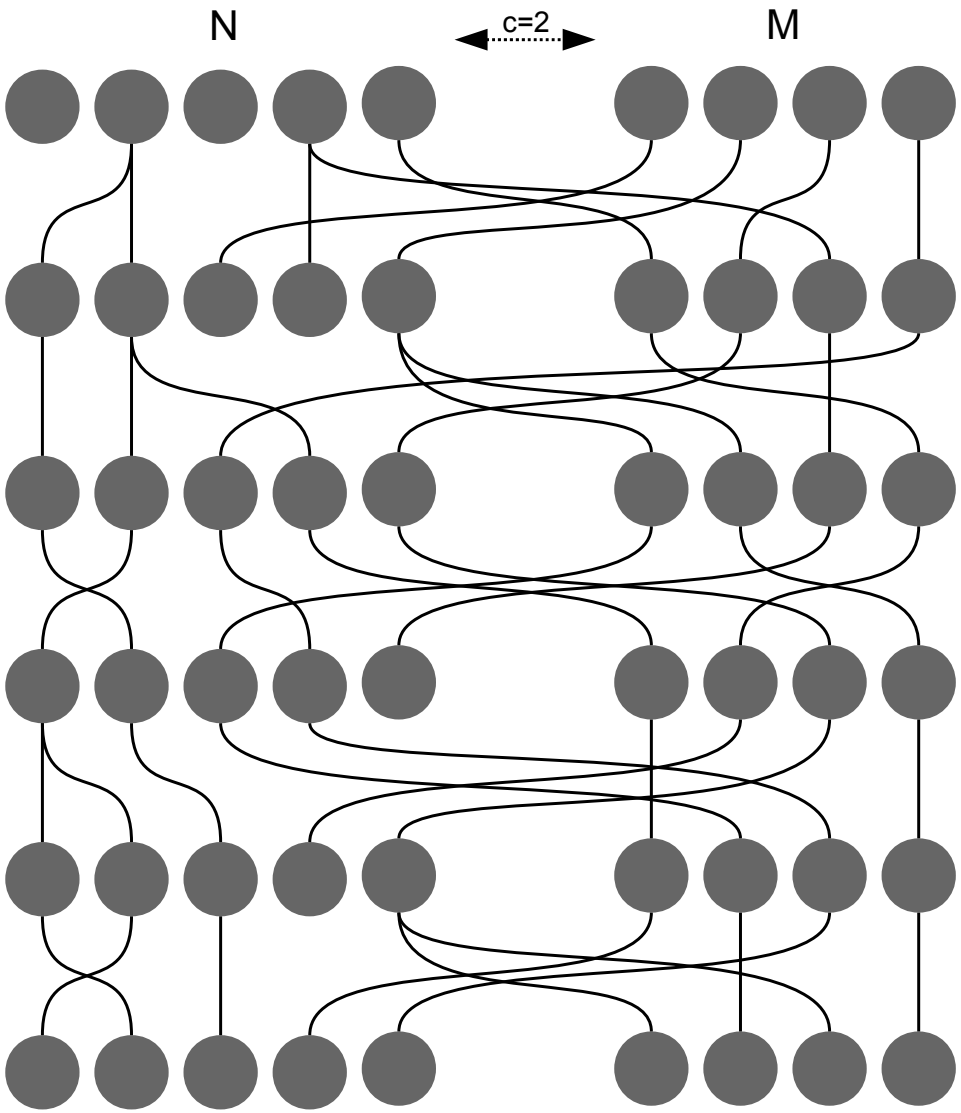
The Wright-Fisher model with geometric seed-bank



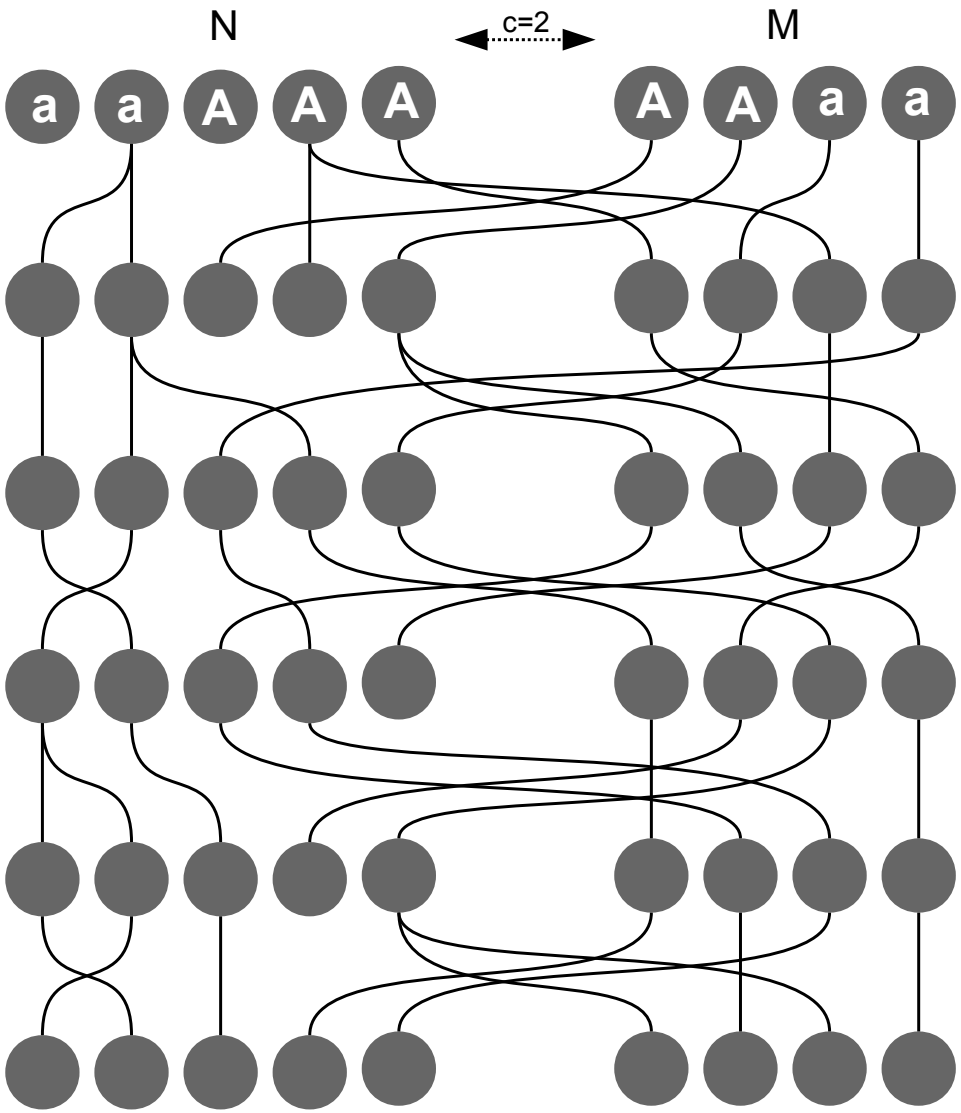
The Wright-Fisher model with geometric seed-bank



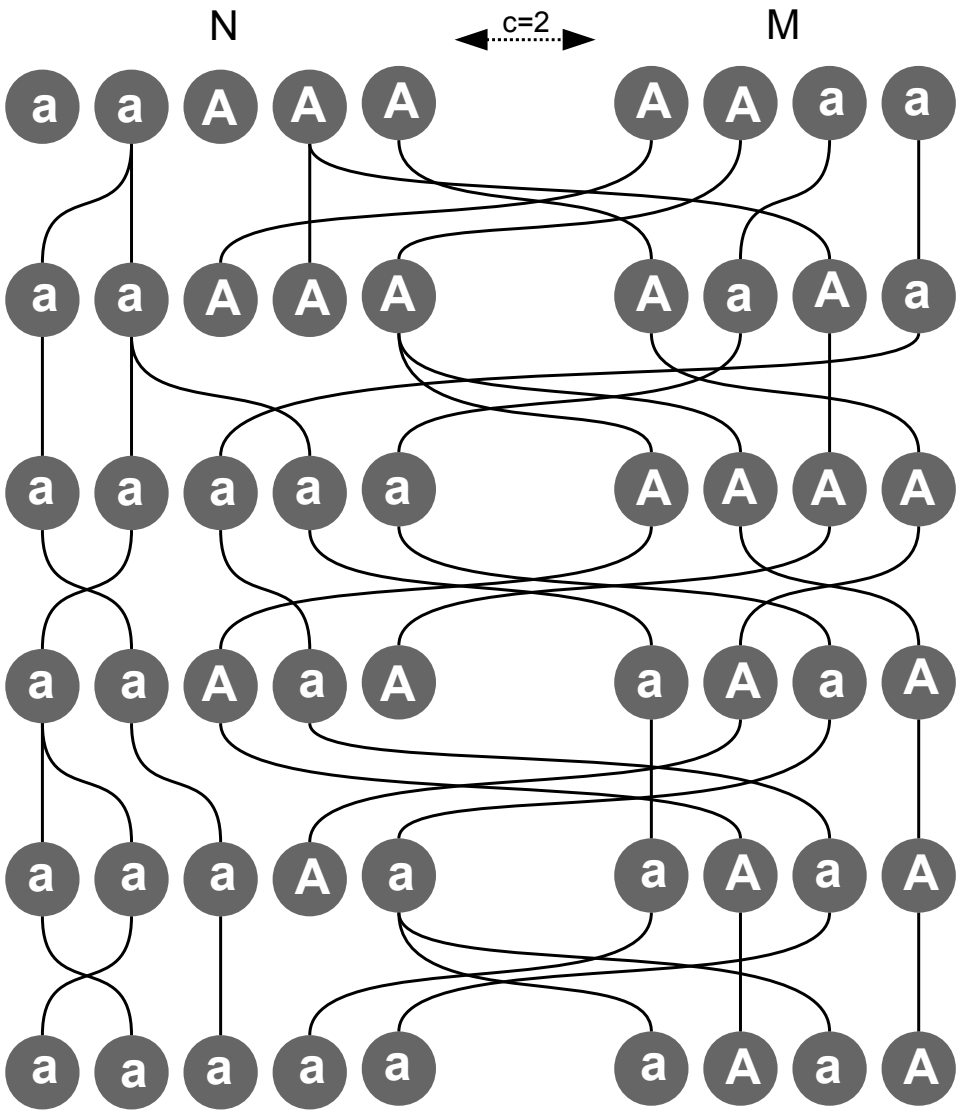
The Wright-Fisher model with geometric seed-bank



The Wright-Fisher model with geometric seed-bank



The Wright-Fisher model with geometric seed-bank



Formal definition, notation

Definition 1.1 (Wright-Fisher model with geometric seed-bank component)

Fix pop.-size $N \in \mathbb{N}$, seed-bank size M, genetic type space E and parameters δ, ε as before. Given initial type configurations $\xi_0 \in E^N$ and $\eta_0 \in E^M$, let

$$\xi_k := \big(\xi_k(i), i \in \{1, \ldots, N\}\big), \quad k \in \mathbb{N},$$

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

$$\eta_k := \big(\eta_k(j), j \in \{1, \ldots, M\}\big), \quad k \in \mathbb{N},$$

be 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*.

Age structure in seed bank

Note that the time that a seed stays in the seed bank is iid geometric with success parameter δ.

We will later let ε, δ (and M) scale with N, and in particular assume that $\varepsilon = \varepsilon(N) = c/N$ and $N = K \cdot M(N)$ for constant $c, K \in (0, \infty)$.

Then, the seed-bank age distribution is geometric with parameter cK/N, and in particular the average time spent in the dormant state is $N/cK \in O(N)$.

Age structure in seed bank

Note that the time that a seed stays in the seed bank is iid geometric with success parameter δ.

We will later let ε, δ (and M) scale with N, and in particular assume that $\varepsilon = \varepsilon(N) = c/N$ and $N = K \cdot M(N)$ for constant $c, K \in (0, \infty)$.

Then, the seed-bank age distribution is geometric with parameter cK/N, and in particular the average time spent in the dormant state is $N/cK \in O(N)$.

Age structure in seed bank

Note that the time that a seed stays in the seed bank is iid geometric with success parameter δ.

We will later let ε, δ (and M) scale with N, and in particular assume that $\varepsilon = \varepsilon(N) = c/N$ and $N = K \cdot M(N)$ for constant $c, K \in (0, \infty)$.

Then, the seed-bank age distribution is geometric with parameter cK/N, and in particular the average time spent in the dormant state is $N/cK \in O(N)$.

Frequency chains of the geometric seed-bank model

We now specialise again to the bi-allelic case $E = \{a, A\}$.

Definition 1.2 (Frequency chains)

Let

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

The pair forms a discrete-time Markov chain taking values in $I^N \times I^M$, where

$$I^N = \left\{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\right\} \quad \text{and} \quad I^M = \left\{0, \frac{1}{M}, \frac{2}{M}, \ldots, 1\right\}.$$

Let $\mathbb{P}_{x,y}$ be the law of (X^N, Y^M) with initial frequencies x, y.

The limiting generator of the allele frequency processes

For $c, K \in (0, \infty)$ assume

$$\varepsilon = \varepsilon(N) = \frac{c}{N}, \quad M = M(N) = \frac{N}{K}, \quad \text{and} \quad \delta = \delta(N) = \frac{c}{M(N)} = \frac{cK}{N}.$$

Proposition 1.3

For the above parameter choices, and suitable test functions f, consider the discrete generator A^N of $(X_k^N, Y_k^M)_{k\in\mathbb{N}}$:

$$A^N f(x,y) := N\,\mathbb{E}_{x,y}\Big[f\big(X_1^N, Y_1^M\big) - f(x,y)\Big].$$

Then, we have

$$\begin{aligned} Af(x,y) &:= \lim_{N\to\infty} A^N f(x,y) \\ &= c(y-x)\frac{\partial f}{\partial x}(x,y) + cK(x-y)\frac{\partial f}{\partial y}(x,y) + \frac{1}{2}x(1-x)\frac{\partial f^2}{\partial x^2}(x,y). \end{aligned}$$

The scaling-limit of the allele frequency process

The result follows from the usual Taylor expansion of $A^N f$ about (x, y) (lengthy details omitted). We arrive at the following

Corollary 1.4 (Wright-Fisher diffusion with seed bank)

Under the conditions of Proposition 1.3,

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

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

$$\begin{aligned} dX_t &= c(Y_t - X_t)dt + \sqrt{X_t(1 - X_t)}dB_t, \\ dY_t &= cK(X_t - Y_t)dt, \end{aligned} \tag{2}$$

with $X_0 = x$, $Y_0 = y$.

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.

Such dual processes are often extremely useful in 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.

The dual of the seed-bank frequency process

Definition 1.5

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 started in $(N_0, M_0) \in \mathbb{N}^2$ with transitions

$$\begin{aligned} (n,m) &\mapsto (n-1, m+1) \quad \text{at rate} \quad cn \\ (n,m) &\mapsto (n+1, m-1) \quad \text{at rate} \quad cKm \\ (n,m) &\mapsto (n-1, m) \quad \text{at rate} \quad \binom{n}{2} \end{aligned}$$

Denote by $\mathbb{P}^{n,m}$ the distribution of $(N_t, M_t)_{t\geq 0}$ if started in $(N_0, M_0) = (n, m)$, and denote the corresponding expected value by $\mathbb{E}^{n,m}$.

Moment duality

It is easy to see that *eventually*, $N_t + M_t = 1$ (as $t \to \infty$), since the sum $M + N$ can be dominated by a pure death process.

Moreover, it is standard to show that $(N_t, M_t)_{t \geq 0}$ is the *moment dual* of $(X_t, Y_t)_{t \geq 0}$.

Theorem 1.6

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

$$\mathbb{E}_{x,y}\left[X_t^n Y_t^m\right] = \mathbb{E}^{n,m}\left[x^{N_t} y^{M_t}\right].$$

We aim to exploit this duality in order to learn something about the long-term behaviour of our system.

The long-time behaviour (in law)

The long-term behaviour of our system (2) is not obvious. While a classical Wright Fisher diffusion $\{Z_t\}$, given by

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

will get absorbed at the boundaries a.s. after finite time (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.

Obviously, $(0, 0)$ and $(1, 1)$ are absorbing states for the system (2).

The long-time behaviour (in law)

The long-term behaviour of our system (2) is not obvious. While a classical Wright Fisher diffusion $\{Z_t\}$, given by

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

will get absorbed at the boundaries a.s. after finite time (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.

Obviously, $(0, 0)$ and $(1, 1)$ are absorbing states for the system (2).

The long-time behaviour (in law)

One can compute its 'fixation probability' as $t \to \infty$, in a suitable sense (in law). We prepare this with a moment computation.

Proposition 1.7

All mixed moments of $(X_t, Y_t)_{t\geq 0}$ solving (2) converge to the same finite limit depending only on x, y, K. More precisely, for each fixed $n, m \in \mathbb{N}$, we have

$$\lim_{t\to\infty} \mathbb{E}_{x,y}[X_t^n Y_t^m] = \lim_{t\to\infty} \mathbb{E}^{n,m}\left[x^{N_t} y^{M_t}\right] = \frac{y + xK}{1 + K}. \tag{3}$$

The long-time behavior (in law), II

Corollary 1.8 (Fixation in law)

Given c, K and $(X_0, Y_0) = (x, y) \in [0,1]^2$, we have that

$$\lim_{t\to\infty} \mathcal{L}(X_t, Y_t) = \frac{y + xK}{1+K}\delta_{(1,1)} + \frac{1 + (1-x)K - y}{1+K}\delta_{(0,0)}.$$

Note that this is in line with the classical results for the Wright-Fisher diffusion:

As $K \to \infty$ (that is, 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.

The long-time behavior (in law), II

Corollary 1.8 (Fixation in law)

Given c, K and $(X_0, Y_0) = (x, y) \in [0,1]^2$, we have that

$$\lim_{t\to\infty} \mathcal{L}(X_t, Y_t) = \frac{y + xK}{1+K}\delta_{(1,1)} + \frac{1 + (1-x)K - y}{1+K}\delta_{(0,0)}.$$

Note that this is in line with the classical results for the Wright-Fisher diffusion:

As $K \to \infty$ (that is, 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.

The long-time behavior (in law), II

Corollary 1.8 (Fixation in law)

Given c, K and $(X_0, Y_0) = (x, y) \in [0,1]^2$, we have that

$$\lim_{t\to\infty} \mathcal{L}(X_t, Y_t) = \frac{y + xK}{1+K}\delta_{(1,1)} + \frac{1 + (1-x)K - y}{1+K}\delta_{(0,0)}.$$

Note that this is in line with the classical results for the Wright-Fisher diffusion:

As $K \to \infty$ (that is, 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.

Almost sure behaviour?

Note that the above result does *not* fully explain the pathwise/almost-sure picture.

Indeed, absorption will *not* happen in finite time, since the dual *block counting process*, started from an infinite initial state, *does not come down from infinity*, which means that the total (infinite) population does not have a most-recent common ancestor (we will see this later).

Thus, initial genetic variability in a hypothetical infinite population would never be completely lost.

Almost sure behaviour?

Note that the above result does *not* fully explain the pathwise/almost-sure picture.

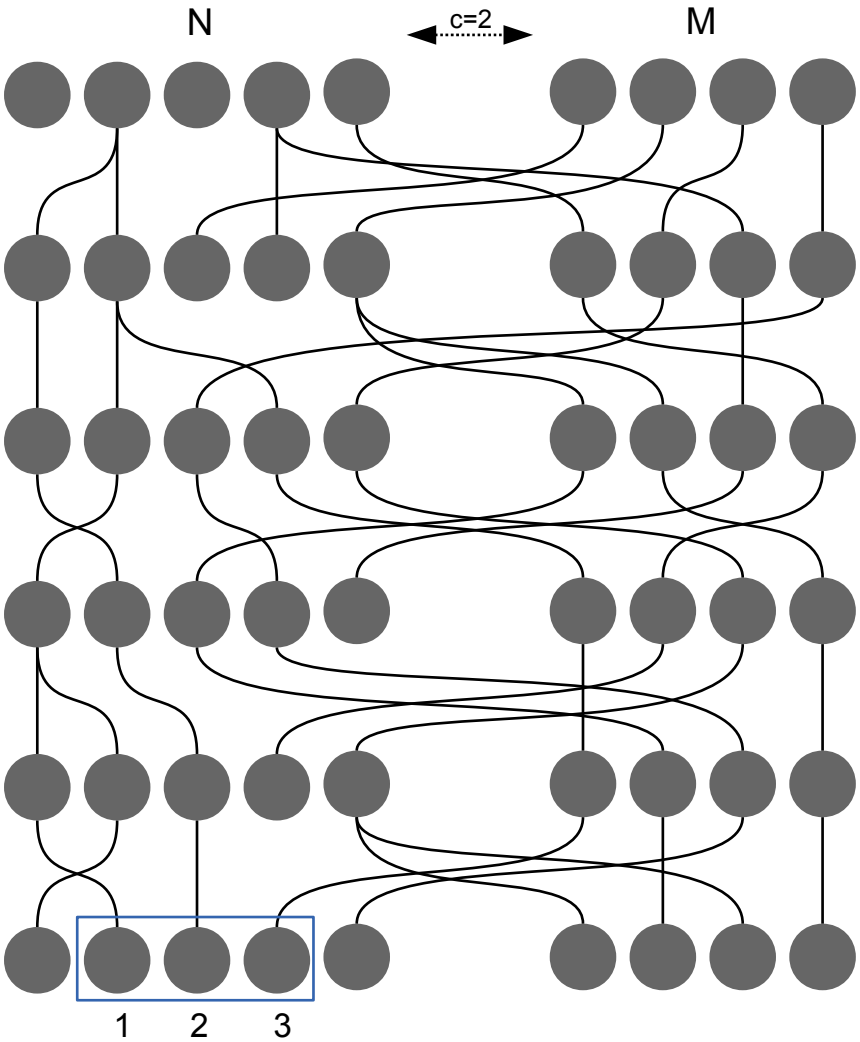
Indeed, absorption will *not* happen in finite time, since the dual *block counting process*, started from an infinite initial state, *does not come down from infinity*, which means that the total (infinite) population does not have a most-recent common ancestor (we will see this later).

Thus, initial genetic variability in a hypothetical infinite population would never be completely lost.

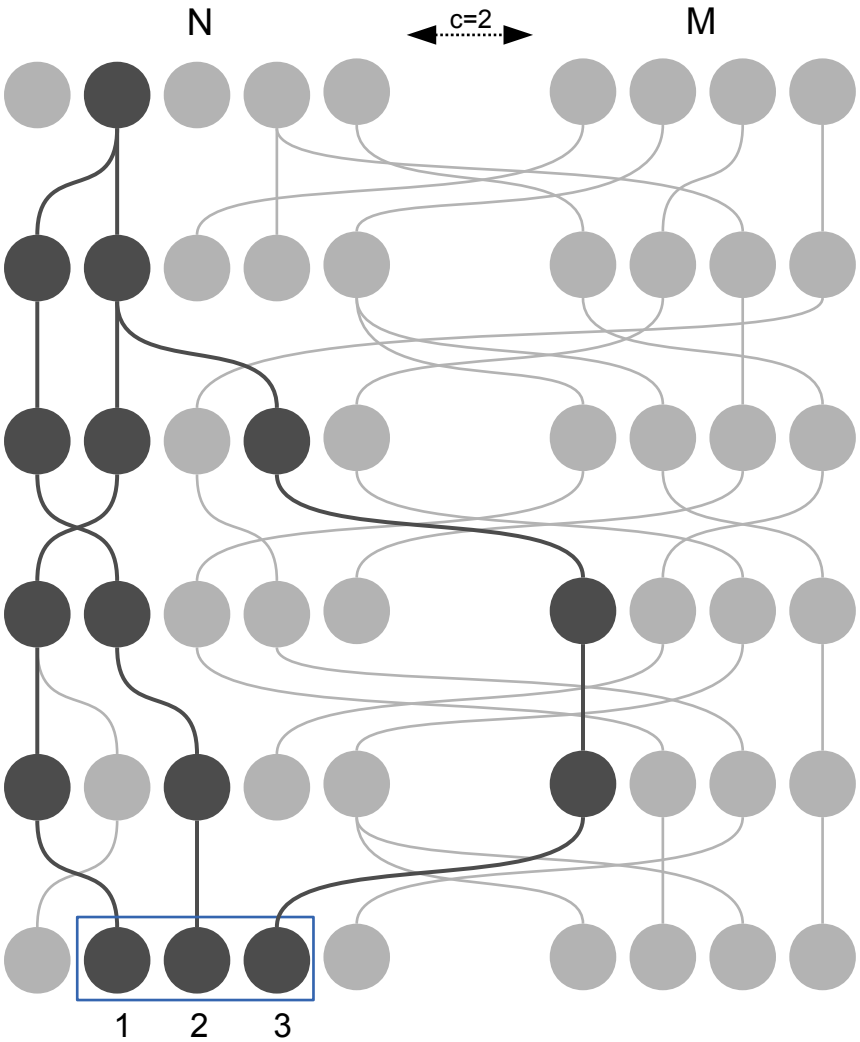
The genealogy of a sample

In view of the form of the block counting process, it is now easy to guess the stochastic process describing the limiting ancestral process of a sample taken from the Wright-Fisher model with geometric seed-bank component.

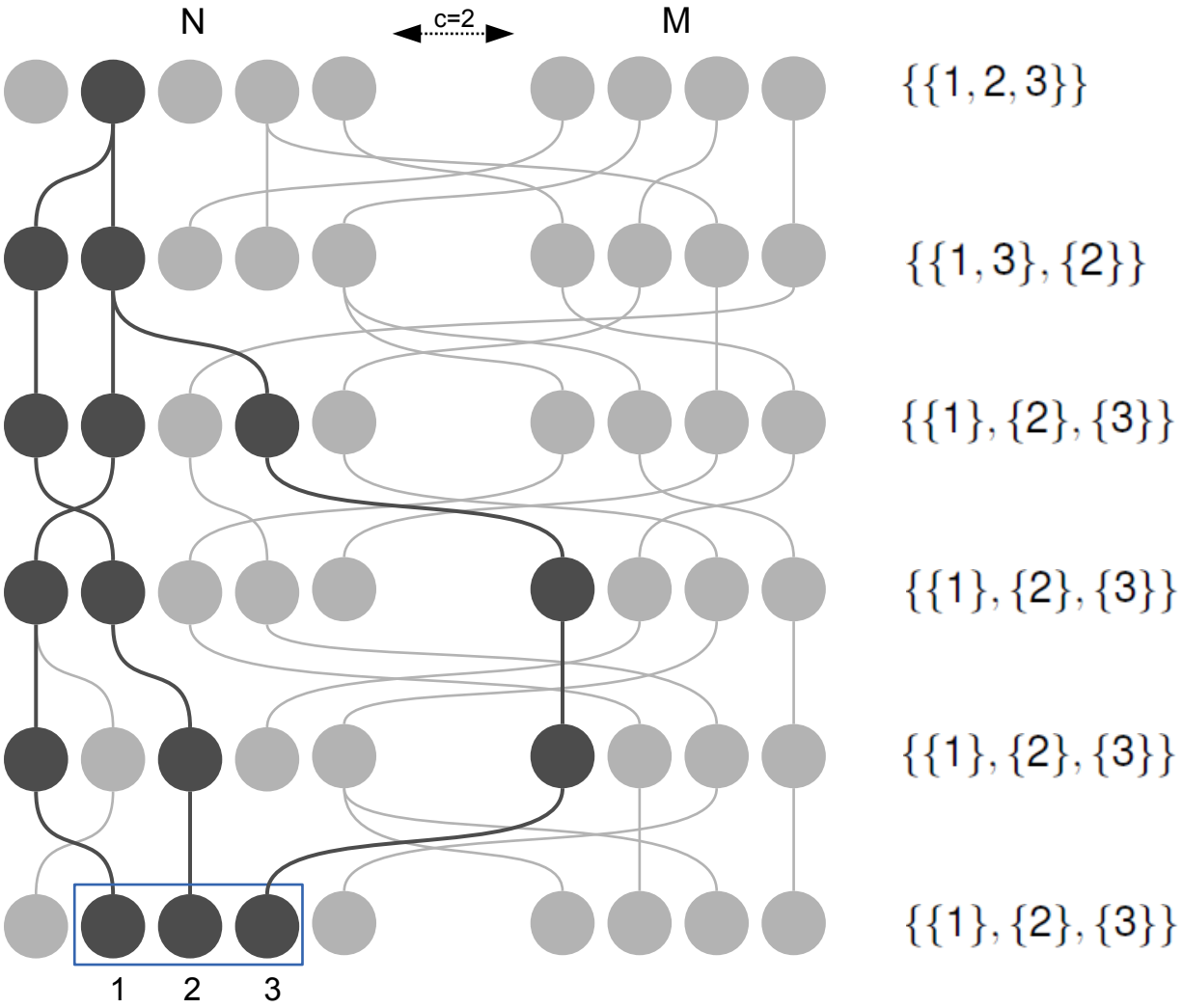
The genealogy of a sample



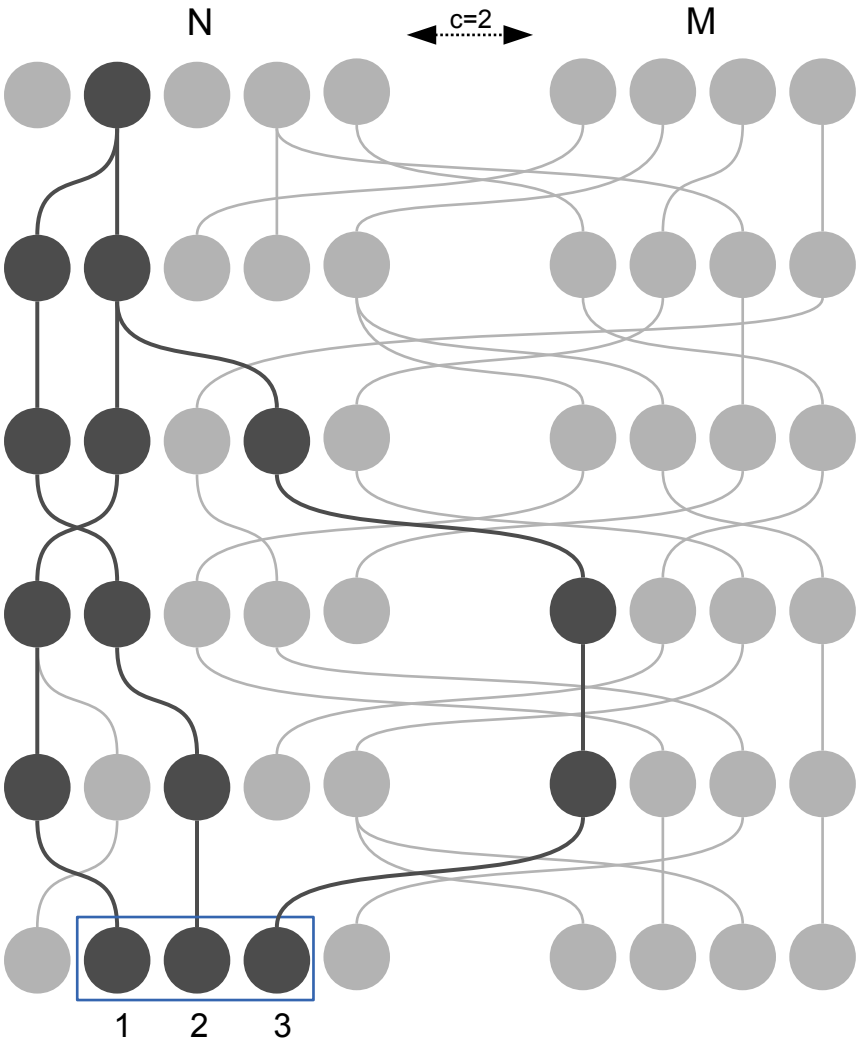
The genealogy of a sample



The genealogy of a sample



The genealogy of a sample



$\{\{1,2,3\}^p\}$

$\{\{1,3\}^p,\{2\}^p\}$

$\{\{1\}^p,\{2\}^p,\{3\}^p\}$

$\{\{1\}^p,\{2\}^p,\{3\}^s\}$

$\{\{1\}^p,\{2\}^p,\{3\}^s\}$

$\{\{1\}^p,\{2\}^p,\{3\}^p\}$

The genealogy of a sample and marked partitions

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

$$\mathcal{P}_k^{\{p,s\}} = \left\{ \pi \times \{s,p\}^{|\pi|} : \pi \in \mathcal{P}_k \right\}.$$

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

$$\pi = \left\{ \{1,3\}^p \{2\}^s \{4,5\}^p \right\}.$$

The genealogy of a sample and partitions with flags, II

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

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

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

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

The seed bank coalescent

Definition 1.9

For $k \geq 1$ and $c, K \in (0, \infty)$ we define the k-*seed bank coalescent* $(\Pi_t^{(k)})_{t \geq 0}$ with seed-bank intensity c and seed-bank size $1/K$ to be the continuous time pure-jump Markov process with values in $\mathcal{P}_k^{\{p,s\}}$, with transitions:

$$\begin{aligned}
\pi \to \pi' \quad & \text{at rate } 1 \text{ if } \pi \succ \pi', \\
\pi \to \pi' \quad & \text{at rate } c \text{ if } \pi \bowtie \pi' \text{ and one } p \text{ is replaced by one } s, \\
\pi \to \pi' \quad & \text{at rate } cK \text{ if } \pi \bowtie \pi' \text{ and one } s \text{ is replaced by one } p.
\end{aligned}$$

If $c = K = 1$, we speak of the *standard (k-) seed bank coalescent*.

The seed bank coalescent - Illustration

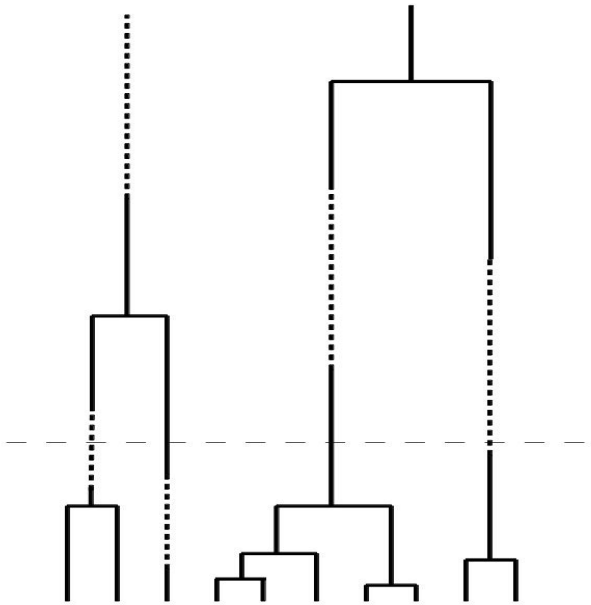


Figure: A possible realisation of the standard 10-seed bank coalescent. Dotted lines indicate 'inactive lineages' (carrying an s-flag, which are prohibited from merging). At the time marked with the dotted horizontal line the process is in state $\{\{1,2\}^s\{3\}^p\{4,5,6,7,8\}^p\{9,10\}^s\}$.

The seed bank coalescent as scaling limit

- 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 in the classical Wright-Fisher model.
- Indeed, consider the genealogy of a sample of n ($\ll N$) individuals, sampled from present generation 0. Denote by $\Pi_i^{(N,n)} \in \mathcal{P}_n^{\{p,s\}}$ the partition at generation $-i$, where two individuals belong to the same block of $\Pi_i^{(N,n)}$ if and only if their ancestral lines have met before generation $-i$.
- The flag s or p indicates whether the ancestor in generation $-i$ is a plant or a seed.

The seed bank coalescent as scaling limit

- 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 in the classical Wright-Fisher model.
- Indeed, consider the genealogy of a sample of n ($\ll N$) individuals, sampled from present generation 0. Denote by $\Pi_i^{(N,n)} \in \mathcal{P}_n^{\{p,s\}}$ the partition at generation $-i$, where two individuals belong to the same block of $\Pi_i^{(N,n)}$ if and only if their ancestral lines have met before generation $-i$.
- The flag s or p indicates whether the ancestor in generation $-i$ is a plant or a seed.

The seed bank coalescent as scaling limit

- 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 in the classical Wright-Fisher model.
- Indeed, consider the genealogy of a sample of n ($\ll N$) individuals, sampled from present generation 0. Denote by $\Pi_i^{(N,n)} \in \mathcal{P}_n^{\{p,s\}}$ the partition at generation $-i$, where two individuals belong to the same block of $\Pi_i^{(N,n)}$ if and only if their ancestral lines have met before generation $-i$.
- The flag s or p indicates whether the ancestor in generation $-i$ is a plant or a seed.

The seed bank coalescent as scaling limit

Standard arguments now give the following:

Corollary 1.10

Under the assumptions of Proposition 1.3, $(\Pi^{(N,n)}_{\lfloor Nt \rfloor})$ *converges weakly as* $N \to \infty$ *to the seed-bank coalescent* $(\Pi^{(n)}_t)$ *started with n plants.*

Properties of the seed bank coalescent

It is not surprising that the seed bank coalescent behaves very differently from a classical Kingman coalescent.

We illustrate this with two examples.

Coming down from infinity

The notion of *coming down from infinity* was introduced by [Pitman 1999] and [Schweinsberg 2000]. 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. for all $t > 0$ a.s.).

Coming down from infinity

Theorem 1.11 ([BGCKWB15])

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, that is, for each infinite starting configuration (N_0, M_0) with $N_0 = \infty$,

$$\mathbb{P}\big\{N_t + M_t = \infty, \text{ for all } t > 0\big\} = 1.$$

Of course, this has to do with lineages immediately 'escaping' into the seed-bank.

Time to the most recent common ancestor

The seed bank causes a significant delay in the time to the most recent common ancestor.

Definition 1.12

Let $k \in \mathbb{N} \cup \{\infty\}$. We define the *time to the most recent common ancestor* of a sample of n plants to be

$$T_{MRCA}[n] = \inf\{t > 0 : |\Pi_t^{(n)}| = 1 \text{ given that } \Pi_0 = \{\{1\}^p ... \{n\}^p\}\},$$

or equivalently

$$T_{MRCA}[n] = \inf\{t > 0 : (N_t, M_t) = (1, 0) \text{ given that } (N_0, M_0) = (n, 0)\}$$

Time to the most recent common ancestor

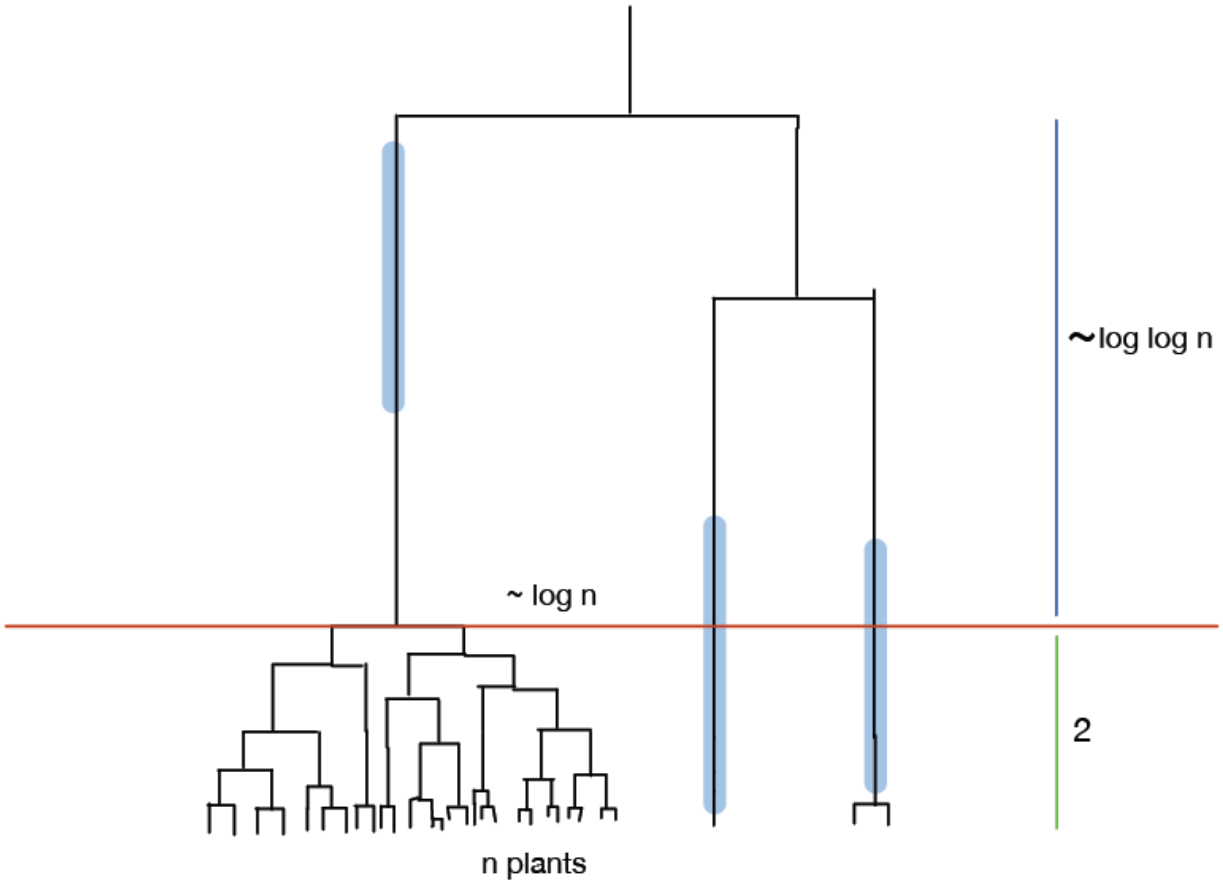
Theorem 1.13

For all $c, K \in (0, \infty)$, the seed bank coalescent satisfies

$$\mathbb{E}\big[T_{MRCA}[n]\big] \asymp \log\log n. \tag{4}$$

This should be compared with a result for the *Bolthausen-Sznitman coalescent* in [Goldschmidt & Martin 2005], which exhibits the same time scale.

Time to the most recent common ancestor



Extensions of the model

The seed-bank model can be extended to include:

- *Mutation*, both in the active and the dormant population. The usual scaling leads to a seed bank coalescent with mutation, where *Poisson mutations* appear with rate $\theta_1/2$ on the active lines, and with rate $\theta_2/2$ on the dormant lines.
- *Mortality* in the seed bank: Assume d/N is the death probability per individual per generation in the seed bank, and assume that vacant slots due to seed deaths are filled by additional offspring from the active population. This leads to a model with an *'effective' seed bank parameter*

$$\tilde{K} = \frac{c+d}{c} K.$$

- There is a *Moran-model* formulation, and a corresponding *lookdown construction* (work in progress with [Ch. Horvath, TUB]).

Extensions of the model

The seed-bank model can be extended to include:

- *Mutation*, both in the active and the dormant population. The usual scaling leads to a seed bank coalescent with mutation, where *Poisson mutations* appear with rate $\theta_1/2$ on the active lines, and with rate $\theta_2/2$ on the dormant lines.
- *Mortality* in the seed bank: Assume d/N is the death probability per individual per generation in the seed bank, and assume that vacant slots due to seed deaths are filled by additional offspring from the active population. This leads to a model with an *'effective' seed bank parameter*

$$\tilde{K} = \frac{c+d}{c} K.$$

- There is a *Moran-model* formulation, and a corresponding *lookdown construction* (work in progress with [Ch. Horvath, TUB]).

Extensions of the model

The seed-bank model can be extended to include:

- *Mutation*, both in the active and the dormant population. The usual scaling leads to a seed bank coalescent with mutation, where *Poisson mutations* appear with rate $\theta_1/2$ on the active lines, and with rate $\theta_2/2$ on the dormant lines.
- *Mortality* in the seed bank: Assume d/N is the death probability per individual per generation in the seed bank, and assume that vacant slots due to seed deaths are filled by additional offspring from the active population. This leads to a model with an *'effective' seed bank parameter*

$$\tilde{K} = \frac{c+d}{c} K.$$

- There is a *Moran-model* formulation, and a corresponding *lookdown construction* (work in progress with [Ch. Horvath, TUB]).

The seed bank coalescent with mutation

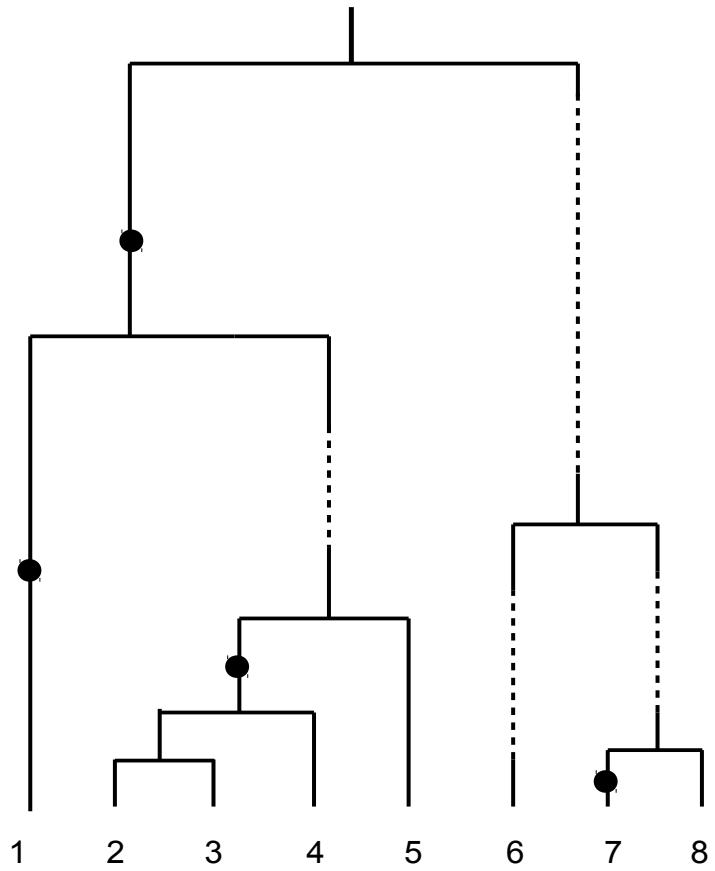


Figure: A possible realisation of the standard 10-seed bank coalescent with mutation (only on active lineages, that is, $\theta_2 = 0$).

Universality

The seed bank coalescent can be expected to arise as universal scaling limit under the following assumptions:

- The active population, without seeds, is in the domain of attraction of the Kingamn coalescent, say, with time-scaling N (or, more generally $1/c_N$ where c_N is the probability that two individuals in a Cannings model share a common ancestor in the previous generation).
- The seed bank is of comparable size as the active population ($O(N)$), up to fluctuations of smaller order.
- Initiation and resuscitation happen at a probability / rate of order $O(1/N)$. That means that dormancy times are on the order of the active population size.
- Mutations appear at probability / rate $O(1/N)$.

Universality

The seed bank coalescent can be expected to arise as universal scaling limit under the following assumptions:

- The active population, without seeds, is in the domain of attraction of the Kingamn coalescent, say, with time-scaling N (or, more generally $1/c_N$ where c_N is the probability that two individuals in a Cannings model share a common ancestor in the previous generation).
- The seed bank is of comparable size as the active population ($O(N)$), up to fluctuations of smaller order.
- Initiation and resuscitation happen at a probability / rate of order $O(1/N)$. That means that dormancy times are on the order of the active population size.
- Mutations appear at probability / rate $O(1/N)$.

Universality

The seed bank coalescent can be expected to arise as universal scaling limit under the following assumptions:

- The active population, without seeds, is in the domain of attraction of the Kingamn coalescent, say, with time-scaling N (or, more generally $1/c_N$ where c_N is the probability that two individuals in a Cannings model share a common ancestor in the previous generation).
- The seed bank is of comparable size as the active population ($O(N)$), up to fluctuations of smaller order.
- Initiation and resuscitation happen at a probability / rate of order $O(1/N)$. That means that dormancy times are on the order of the active population size.
- Mutations appear at probability / rate $O(1/N)$.

Universality

The seed bank coalescent can be expected to arise as universal scaling limit under the following assumptions:

- The active population, without seeds, is in the domain of attraction of the Kingamn coalescent, say, with time-scaling N (or, more generally $1/c_N$ where c_N is the probability that two individuals in a Cannings model share a common ancestor in the previous generation).
- The seed bank is of comparable size as the active population ($O(N)$), up to fluctuations of smaller order.
- Initiation and resuscitation happen at a probability / rate of order $O(1/N)$. That means that dormancy times are on the order of the active population size.
- Mutations appear at probability / rate $O(1/N)$.

Relation to other models

The seed bank diffusion is related to two other interesting stochastic systems (work in progress with [E. Buzzoni, A. González & A. Etheridge]):

- Seed bank diffusion scaling limit (with mutation) can be reformulated in terms of a *stochastic delay differential equation*.
- Related to the two island model (where coalescences are completely blocked in one island) and the *structured coalescent* [Herbots 1997].

Characterization of stationary distribution?

Relation to other models

The seed bank diffusion is related to two other interesting stochastic systems (work in progress with [E. Buzzoni, A. González & A. Etheridge]):

- Seed bank diffusion scaling limit (with mutation) can be reformulated in terms of a *stochastic delay differential equation*.
- Related to the two island model (where coalescences are completely blocked in one island) and the *structured coalescent* [Herbots 1997].

Characterization of stationary distribution?

Relation to other models

The seed bank diffusion is related to two other interesting stochastic systems (work in progress with [E. Buzzoni, A. González & A. Etheridge]):

- Seed bank diffusion scaling limit (with mutation) can be reformulated in terms of a *stochastic delay differential equation*.
- Related to the two island model (where coalescences are completely blocked in one island) and the *structured coalescent* [Herbots 1997].

Characterization of stationary distribution?

Seed banks in bacterial communities

The universality assumptions *may* be fitting relatively well to those stated in [Jones & Lennon 2010, Lennon & Jones 2011] investigating bacterial communities:

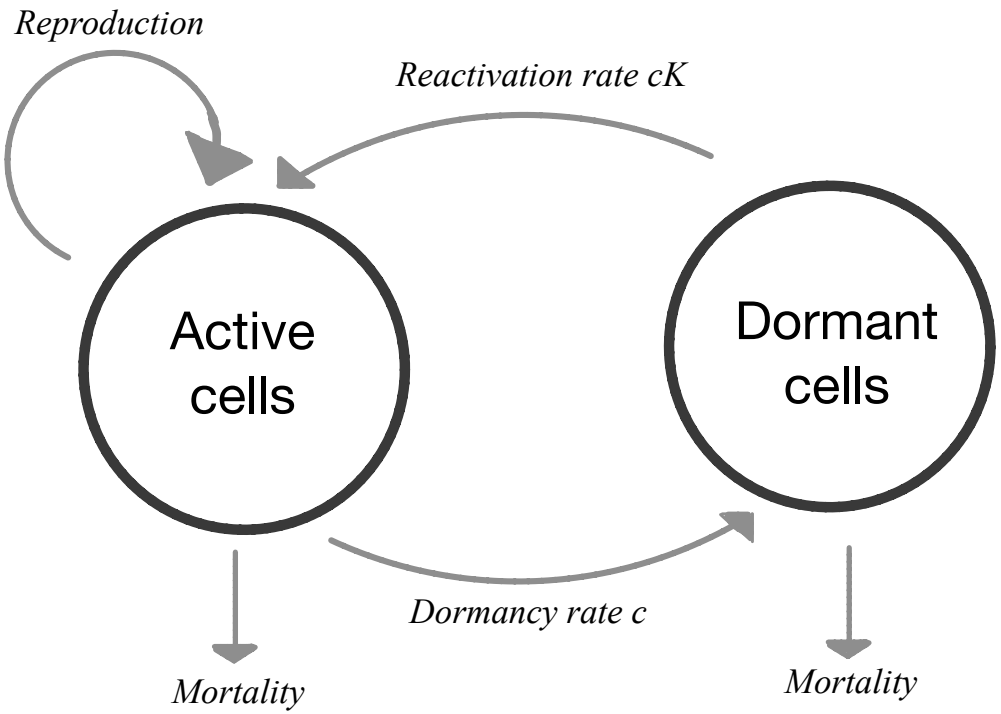


Figure: Initiaton and resuscitation, à la [Lennon & Jones 2011]

Genetic variability under the seed bank coalescent

The seed bank coalescent allows the derivation of recursions for expected values (and variances, covariances...) of

- *tree quantities*, such as TMRCA, total tree length, external branch lengths, ...
- and thus classical *population genetic quantities* such as segregating sites, pairwise differences, singletons, SFS, etc...

Also, sampling formulas can be attacked, at least via recursions.

Search for explicit (limiting) distributions poses interesting mathematical challenges.

Genetic variability under the seed bank coalescent

The seed bank coalescent allows the derivation of recursions for expected values (and variances, covariances...) of

- *tree quantities*, such as TMRCA, total tree length, external branch lengths, ...
- and thus classical *population genetic quantities* such as segregating sites, pairwise differences, singletons, SFS, etc...

Also, sampling formulas can be attacked, at least via recursions.

Search for explicit (limiting) distributions poses interesting mathematical challenges.

Expected time to MRCA - exact values

We already know that $\mathbb{E}_{n,0}[T_{\mathrm{MRCA}}]$ for the seed bank coalescent, if started in a sample of active individuals of size n, is $O(\log\log n)$.

However, this does not give precise information for the exact absolute value, in particular for 'small to medium' n. Instead, we can derive a recursion:

For $n, m \in \mathbb{N}_0$ let

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

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

Expected time to MRCA - exact values

We already know that $\mathbb{E}_{n,0}[T_{\mathrm{MRCA}}]$ for the seed bank coalescent, if started in a sample of active individuals of size n, is $O(\log\log n)$.

However, this does not give precise information for the exact absolute value, in particular for 'small to medium' n. Instead, we can derive a recursion:

For $n, m \in \mathbb{N}_0$ let

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

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

Expected time to MRCA - exact values

We already know that $\mathbb{E}_{n,0}[T_{\mathrm{MRCA}}]$ for the seed bank coalescent, if started in a sample of active individuals of size n, is $O(\log\log n)$.

However, this does not give precise information for the exact absolute value, in particular for 'small to medium' n. Instead, we can derive a recursion:

For $n, m \in \mathbb{N}_0$ let

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

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

Recursion for the expected time to MRCA

Further, abbreviate

$$\lambda_{n,m} := \binom{n}{2} + cn + cKm,$$

and

$$\alpha_{n,m} := \frac{\binom{n}{2}}{\lambda_{n,m}}, \quad \beta_{n,m} := \frac{cn}{\lambda_{n,m}}, \quad \gamma_{n,m} := \frac{cKm}{\lambda_{n,m}}.$$

Then, conditioning on the first transition event (which is exponential with rate $\lambda_{n,m}$) we get

Proposition 1.14

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

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

Special case $n = 2$

For example, one gets

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

Interestingly, $t_{2,0}$ is independent of c, and in particular does *not* converge to 1 (the Kingman case) as $c \to 0$.

This effect is similar to the corresponding behaviour of the structured coalescent if the migration rate goes to 0, cf. [Nath & Griffith 1993] (but there, the factor is 2).

Yet, the Kingman coalescent times are recovered as the relative seed bank size decreases to 0 (i.e. $K \to \infty$).

Special case $n = 2$

For example, one gets

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

Interestingly, $t_{2,0}$ is independent of c, and in particular does *not* converge to 1 (the Kingman case) as $c \to 0$.

This effect is similar to the corresponding behaviour of the structured coalescent if the migration rate goes to 0, cf. [Nath & Griffith 1993] (but there, the factor is 2).

Yet, the Kingman coalescent times are recovered as the relative seed bank size decreases to 0 (i.e. $K \to \infty$).

Special case $n = 2$

For example, one gets

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

Interestingly, $t_{2,0}$ is independent of c, and in particular does *not* converge to 1 (the Kingman case) as $c \to 0$.

This effect is similar to the corresponding behaviour of the structured coalescent if the migration rate goes to 0, cf. [Nath & Griffith 1993] (but there, the factor is 2).

Yet, the Kingman coalescent times are recovered as the relative seed bank size decreases to 0 (i.e. $K \to \infty$).

Some concrete values for T_{MRCA}

$K = 1$

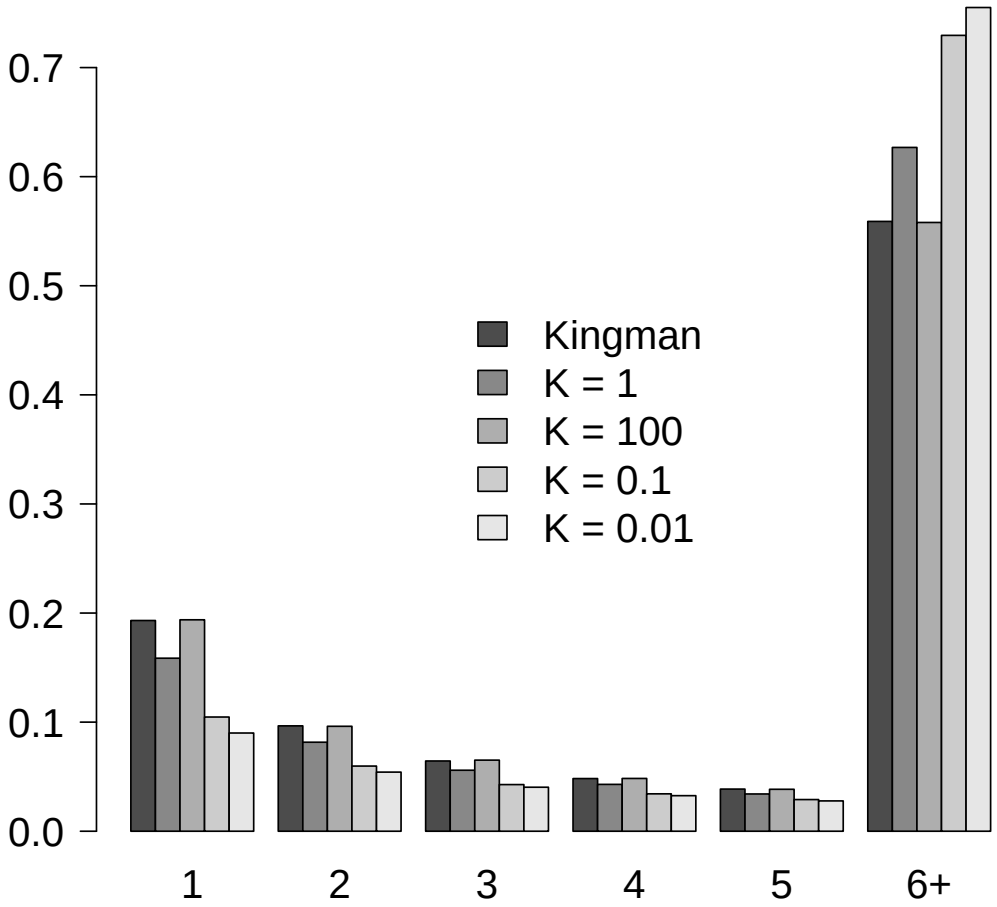
	sample size n		
c	2	10	100
0.01	4	10.21	17.18
0.1	4	9.671	14.97
1	4	8.071	10.02
10	4	7.317	8.221
100	4	7.212	7.954

$K = 100$

	sample size n		
c	2	10	100
0.01	1.02	1.846	2.052
0.1	1.02	1.838	2.026
1	1.02	1.836	2.02
10	1.02	1.836	2.02
100	1.02	1.836	2.02
$K = \infty$	1	1.80	1.98

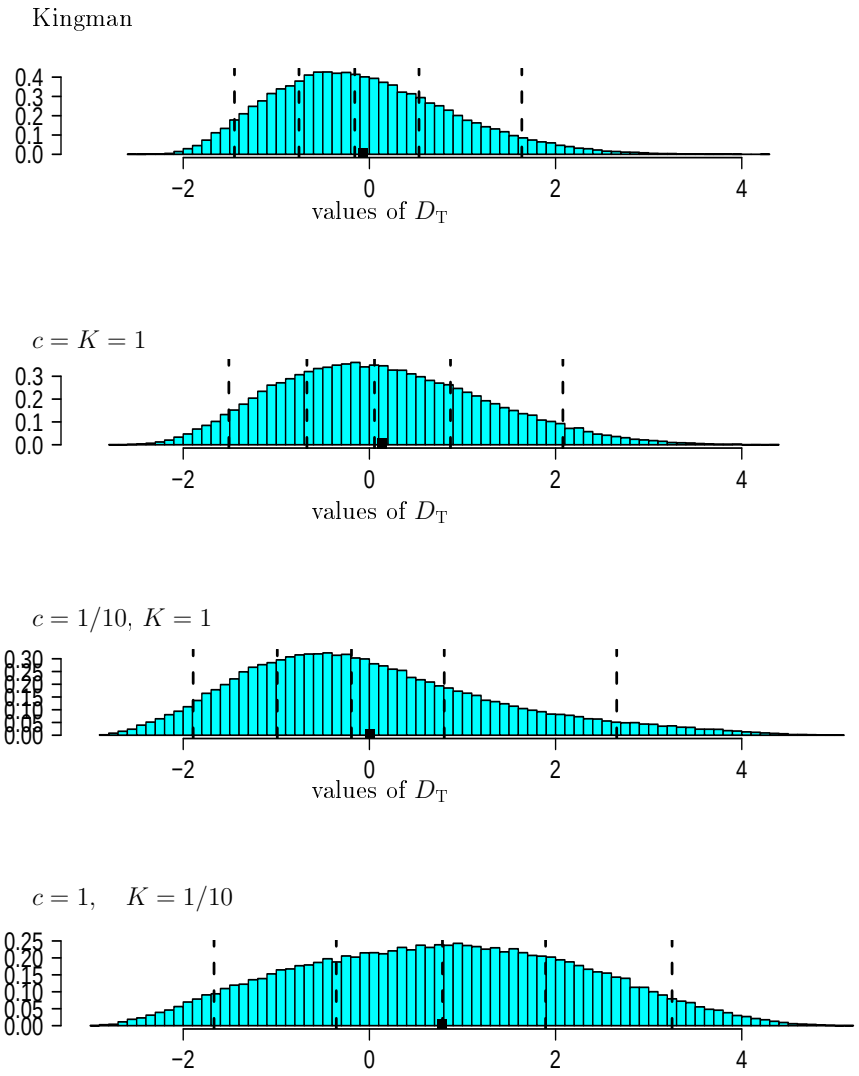
A simulated normalized SFS

A normalized expected site frequency spectrum for various seed bank paramters:



Empirical distribution of Tajima's D

Figure 5: Estimates of the distribution of Tajima's D_{T} (24) with all $n = 100$ sampled lines assumed active, $\theta_1 = 2$, $\theta_2 = 0$. The vertical broken lines are the 5%, 25%, 50%, 75%, 95% quantiles and the black square (■) denotes the mean. The entries are normalised to have unit mass 1. The histograms are drawn on the same horizontal scale. Based on 10^5 replicates.



Many open questions and tasks

- Tree properties of seed bank coalescent...
- Characterization of stationary distribution, properties of system of sdes...
- Modeling extensions: Other evolutionary forces, such as selection, fluctuating population size...
- Modeling extensions: Simultaneous vs. spontaneous switching between dormant and active states
- Derivation of universal limit theorem
- Derivation of lookdown construction
- Statistical analysis: Testing for presence of weak vs. strong seed bank, also for presence of mutation in seed-bank, etc...
- Parameter estimation, efficient simulation, sampling formulas...
- Relation to stochastic delay differential equations, two island models, structured coalescent...
- Cooperation with biologists, application of inference methods to data...

Finally...

... thank you for your attention!

Talk mostly based on:

- Blath, Gonzalèz Casanova, Kurt, Wilke Berenguer: A new coalescent for seed bank models, to appear in *Annals of Applied Probability*, 2015
- Blath, Eldon, Gonzalèz Casanova, Kurt, Wilke Berenguer: Genetic variability under the seed bank coalescent, to appear in *Genetics*, 2015