Coalescent Theory for Seed Bank Models

June 28, 2001

Ingemar Kaj[1], Stephen M. Krone[2], Martin Lascoux[3]

[1]Department of Mathematics, Uppsala University, S-751 06 Uppsala, Sweden. e-mail: ikaj@math.uu.se
[2]Department of Mathematics, University of Idaho, Moscow, ID 83844-1103, USA. e-mail: krone@uidaho.edu
[3]Department of Conservation Biology and Genetics, EBC, Uppsala University, S-752 36 Uppsala, Sweden. e-mail: Martin.Lascoux@ebc.uu.se

Abstract

We study the genealogical structure of samples from a population for which any given generation is made up of direct descendents from several previous generations. These occur in nature when there are seed banks or egg banks allowing an individual to leave offspring several generations in the future. We show how this temporal structure in the reproduction mechanism causes a decrease in the coalescence rate. We also investigate the effects of age-dependent neutral mutations.

Our main result gives weak convergence of the scaled ancestral process, with the usual diffusion scaling, to a coalescent process which is equivalent to a time-changed version of Kingman's coalescent.

Running head: *Seed Bank Coalescents*
Key words: Coalescent, seed bank, age-dependent mutation, urn model, weak convergence
AMS 2000 Subject Classification: Primary 92D10; 60J70; 60F05. Secondary 92D25

1 Introduction

Seed banks undoubtedly play an important evolutionary role, as illustrated for instance by *Linanthus parryae*, a small desert flower which is abundant in the Mojave desert during favorable years. In 1960 Epling *et al.* demonstrated that large fluctuations in population size were due to the presence of a large seed bank, seeds as old as 6 years still being able to germinate. With the notable exception of Templeton and Levin (1979), seed banks (or any equivalent animal storage form such as resting eggs in copepod or water fleas) have received

little attention from theoretical population geneticists, though empirical studies (e.g. Hairston and De Stasio 1988, Levin 1990, Tonsor *et al.* 1993, Alvarez-Buylla *et al.* 1996, McCue and Holtsford 1998, Gomez and Carvalho 2000) have confirmed their evolutionary importance. Apart from obviously securing the survival of populations in unstable environments, seed banks will also slow down the evolutionary rate of some traits by sequestering a substantial fraction of the gene pool from the influence of microevolutionary processes in each generation (Hairston and De Stasio 1988) or lead to an increased accumulation of mutations (Levin 1990).

The joint effects of demographics (seed bank, migration, extinction and recolonization), mutation, recombination and selection on genetic variation can best be studied through coalescent methods (Kingman 1982; Donnelly and Tavaré 1995). The basic idea of the coalescent is to draw a sample of genes from a population and trace back their ancestry, focusing on the times in the past when two or more genes in the sample derive from a common ancestor. Kingman (1982) showed that, with time properly scaled and large population size, this backward process is well described for a classical Wright–Fisher model by a single time-homogeneous Markov chain called the coalescent. Since then the coalescent has been extended to a wide range of demographic and mutation models (see Donnelly 1999; Donnelly and Tavaré 1995; Möhle 2000; Nordborg 2001) and has been shown to be surprisingly robust to departures from the basic Wright–Fisher model. The seed bank can be viewed as a departure from random mating, since it leads to the separation of individuals into different classes. Intuitively, in the case of the seed bank, the rate of coalescence should be slowed down by the structure inherent in the pool of ancestors, as is the case in a geographically structured population: two lineages will "migrate" among generations before they meet in the same one and coalescence can occur.

Seed banks might also affect genetic variation through their effect on mutation. In the absence of selection, the genetic variation in the sample will depend on the rate of coalescence events and on the mutation rate. Aging of seeds leads to an increase in the mutation rate, most of them probably deleterious (Levin 1990). Hence, two copies of the same ancestral gene might have accumulated different numbers and types of mutations depending on the time length they spent in the seed bank.

In the present article we introduce a neutral seed bank model with haploid Wright–Fisher type dynamics, including constant population size. To aid in the analysis, we begin by giving an equivalent urn model for the corresponding genealogical process. Next we investigate the effect of seed banks on the distribution of coalescence times; i.e., we determine the structure of the limiting coalescent tree. We then consider the effect of neutral age-dependent mutations. The main result gives weak convergence of the scaled ancestral process to a time-changed version of Kingman's coalescent. While it is likely that in most seed banks older seeds produce mostly deleterious mutations, this paper does not address non-neutral mutations.

2 The m-step seed bank model and its coalescent

Imagine a population of plants in which each new generation is comprised of N individuals with proportion b_i, $i = 1, \ldots, m$, coming from seeds which were produced i generations ago. Thus seeds are allowed to "overwinter" for up to m generations. We consider this as a population of N haploid genes, updated in discrete time according to modified (neutral) Wright–Fisher dynamics. Each of the N genes in the new generation is obtained, independently of the others, with probability b_i from a seed that was produced i generations ago, $i = 1, \ldots, m$. When a seed is thus chosen from a given generation, it is obtained by random sampling from the appropriate population. So the new generation of N individuals is formed via multinomial sampling from the previous m generations. We call this the *$m-$step seedbank model*. In the classical Wright–Fisher model, $b_1 = 1$, so all individuals in a given generation have parents in the previous generation.

As with the classical Wright–Fisher model, there is an equivalent "backwards-looking" formulation of the model. Namely, each individual in a given generation "chooses" its parent at random from one of the previous m generations according to the above probabilities, this selection being done independently for each individual.

We wish to describe the *ancestry* of a sample of size n from the current generation and from this to derive the corresponding *coalescent process*. To keep track of the ancestral dynamics (and to make this a Markov chain), we must view the state of the population through a time window consisting of m consecutive generations. As we move back in time, we slide the whole window back, one generation at a time. It is perhaps simplest to formulate this in terms of a "space-time urn model" in which ancestors are represented by balls. We begin with some definitions.

A **cell** is a box consisting of N slots and should be thought of as representing the N individuals in a given generation. Cells are labeled $0, 1, 2, \ldots$, to correspond to generation 0 (the current generation), and generations $1, 2, \ldots$ back in time. An **$m-$window** is an ordered collection of m consecutive cells. The **kth $m-$window** consists of cells $(k, k+1, \ldots, k+m-1)$. See Figure 1.

`Figure 1 about here`

We are now ready to describe the dynamics of the urn model. Start with n balls in the 0th $m-$window $(0, 1, \ldots, m-1)$; these n balls correspond to our sample. If we are taking the sample from living plants in the current generation, this would correspond to having all n balls in cell 0. If the sample consists of some living plants and some seeds from previous generations, then the original n balls could be spread out over several of the cells. At any rate, we have some initial configuration of balls in the 0th $m-$window; we represent this by the vector $(X_1(0), X_2(0), \ldots, X_m(0))$, where $X_i(0)$ is the number of balls in cell number $i-1$ (i.e., in the ith cell of the window for the initial configuration). Each ball in a cell will have its own slot since these correspond to distinct individuals in the corresponding generation. Note that there are initially no balls in cells $m, m+1, \ldots$.

To get the state at the next step (step 1), i.e., the configuration of balls in the 1st $m-$window $(1, 2, \ldots, m)$, we start by sliding the original time window to the right by one cell while holding the balls fixed. Thus all of the balls which were in cells $1, \ldots, m-1$ at step 0 will remain in those cells (in fact, in the same slots within those cells), but their relative positions in the new window have shifted to the left by one cell. This corresponds to the seeds in the sample all being one season closer to origination. The balls which were in cell 0 (the original leftmost cell) are no longer in the new window and must be *relocated*, representing the selection of ages of the seeds in the next ancestral generation of the sample. Each of these $X_1(0)$ balls is independently moved to one of the cells $1, \ldots, m$ according to the probabilities $b_1, \ldots, b_m$, respectively. Whenever a relocated ball is put into a new cell, the slot within that cell is chosen uniformly over the N slots in the cell. If a relocated ball lands on an occupied slot in a cell, the two balls coalesce to one and thus the total number of balls (ancestors) decreases by one. Note that this coalescing only happens during relocation, and can be the result of either a relocated ball landing on one of the fixed balls or two relocated balls landing in the same slot of the same cell. Once all of the balls in the previous leftmost cell have been relocated, we will be left with some number of balls $A_N(1) \leq n$.

Now repeat this procedure to get the states at all successive steps. If the process has been defined up to step k, and hence we have some configuration $(X_1(k), \ldots, X_m(k))$ with $X_i(k)$ denoting the number of balls in cell $i+k-1$ (i.e., the ith cell of the kth $m-$window), we get the state at step $k+1$ by sliding the window one cell to the right and relocating the $X_1(k)$ leftmost balls to cells in the new window as before. Note that, since the slot within a cell for each relocated ball is chosen at random over the N slots in that cell, there is no need to keep track of this information when computing the probabilities of coalescence events.

Thus the urn model can be described by a discrete-time Markov chain $\mathbf{X}(k)$, $k = 0, 1, \ldots$, where $\mathbf{X}(k) = (X_1(k), \ldots, X_m(k))$. Let $\mathbf{R}(k+1) = (R_1(k+1), \ldots, R_m(k+1))$ give the numbers of relocated balls landing in the first, second,..., mth cells of the $m-$window at step $k+1$. Then, given $X_1(k)$, this follows a multinomial distribution:

$$\mathbf{R}(k+1) \sim \text{Mult}\,(X_1(k); b_1, \ldots, b_m).$$

In particular, the marginal number sent to the ith cell is binomial:

$$R_i(k+1) \sim \text{Bin}\,(X_1(k), b_i).$$

We remark that these quantities depend implicitly on the population size N.

Let

$$A_N(k) \equiv X_1(k) + \cdots + X_m(k)$$

denote the number of balls (or ancestors) at step k for the seed bank model with population size N. We stop the process when $A_N(k)$ reaches 1, corresponding to the most recent common ancestor of the sample. Clearly, the above urn model is equivalent to the ancestral process for the $m-$step seedbank model. We will refer to the process $\mathbf{X}(k)$ as the **configuration process**. While there are r ancestors, $r = 1, \ldots, n$, the configuration process moves among the states in **level** r:

$$S_r \equiv \{(x_1, \ldots, x_m) : x_1 + \cdots + x_m = r\}.$$

The initial state $\mathbf{X}(0)$ is a point in S_n corresponding to the sample of size n. The configuration process moves among the states in level n for awhile. When a coalescence occurs on level n, we drop down to level $n-1$ if only one coalescence occurs, or down to a lower level if two or more coalescences occur. When the process reaches a new level, it starts moving around among the states of that level until another coalescence occurs. This continues until we reach level 1; i.e., until a common ancestor of the sample is found. We will show that, as $N \to \infty$, the probability of multiple coalescence events goes to zero rapidly enough that they do not appear in the coalescent.

Our main result states that the limit of the time-rescaled ancestral process, $A_N([Nt])$, as $N \to \infty$, is given by Kingman's n-coalescent run on a slower time scale.

Theorem 1 *Let $E = \{1, \ldots, n\}$, and let $A_N(k),\ k = 0, 1, \ldots$ be the ancestral process for the $m-$step seed bank model defined above. Set $\beta_1 = 1/E(B)$, where $E(B) = \sum_{i=1}^{m} ib_i$ is the expected value of the seed bank age distribution $P(B = j) = b_j$. As $N \to \infty$, the process $(A_N([Nt]))_{t\geq 0}$ converges weakly in $D_E[0, \infty)$ to $(A(t))_{t\geq 0}$, where $(A(t))_{t\geq 0}$ is the continuous-time Markov chain with state space E, initial state n, and infinitesimal generator matrix $Q = (q_{ij})_{i,j\in\{1,\ldots,n\}}$ defined by*

$$q_{ii} = -\beta_1^2 \binom{i}{2},\ i = 2, \ldots, n,$$

$$q_{i,i-1} = \beta_1^2 \binom{i}{2},\ i = 2, \ldots, n,$$

and $q_{ij} = 0$, otherwise.

(Here, $D_E[0, \infty)$ is the usual space of right-continuous functions from $[0, \infty)$ to E with left limits; weak convergence is with respect to the Skorohod topology; cf. Ethier and Kurtz (1986). This is the natural setting for weak convergence in coalescent theory.)

We will give a derivation of the theorem in this section. To help the reader see what is really going on, the last part of the derivation will be heuristic. The more formal details of this part of the proof will be provided in the next section. We remark that a given cell can accumulate balls at each step during which it is in the window and not the leftmost cell. The balls in the leftmost cell of the $m-$window at step k represents all ancestral lines from the sample which correspond to birth events k generations in the past. The other balls in the window correspond to ancestral lines which are in the "seed" phase.

To compute the probabilities of various coalescence events, consider a cell containing ℓ "old" balls and suppose r relocated balls land in this cell.

$$\begin{aligned}
\mathbb{P}(\text{no coalescence}) &= (1 - \frac{\ell}{N})(1 - \frac{\ell+1}{N}) \cdots (1 - \frac{\ell+r-1}{N}) \\
&= 1 - \frac{1}{N}(\ell + (\ell+1) + \cdots + (\ell+r-1)) + \mathcal{O}(\frac{1}{N^2}) \\
&= 1 - \frac{1}{N}(2\ell + r - 1)\frac{r}{2} + \mathcal{O}(\frac{1}{N^2}) \\
&= 1 - \frac{1}{N}\left(\ell r + \binom{r}{2}\right) + \mathcal{O}(\frac{1}{N^2}).
\end{aligned}$$

Hence

$$\mathbb{P}(\geq 1 \text{ coalescence}) = \frac{1}{N}\left(\ell r + \binom{r}{2}\right) + \mathcal{O}(\frac{1}{N^2}).$$

Similarly,

$$\mathbb{P}(\geq 2 \text{ coalescences}) = \mathcal{O}(1/N^2).$$

This is true even if the coalescences happen in different cells. Thus, for large N, we should be able to ignore multiple coalescence events in a cell and in the whole window, and, for a given cell,

$$\mathbb{P}(\text{exactly one coalescence}) = \frac{1}{N}\left(\ell r + \binom{r}{2}\right) + \mathcal{O}(\frac{1}{N^2})$$

when we send r relocated balls to a cell with ℓ occupied slots.

Now combine this with the (conditional) multinomial relocation mechanism to know the appropriate number of balls being relocated to cell i. Given $\mathbf{X}(k)$ and $\mathbf{R}(k+1)$, depending on whether there is zero or one coalescence occuring in cell i,

$$X_i(k+1) = \begin{cases} X_{i+1}(k) + R_i(k+1), & \text{if no coalescence} \\ X_{i+1}(k) + R_i(k+1) - 1, & \text{if } R_i(k+1) \geq 1, \text{ one coalescence.} \end{cases}$$

The latter event happens with probability

$$\frac{1}{N}\left(X_{i+1}(k)R_i(k+1) + \binom{R_i(k+1)}{2}\right) + \mathcal{O}(\frac{1}{N^2}).$$

Note that we have ignored events of probability $\mathcal{O}(1/N^2)$.

Thus, putting together the action in all the cells, we have transitions from $\mathbf{X}(k)$ to

$$\mathbf{X}(k+1) = \begin{cases} \sigma\mathbf{X}(k) + \mathbf{R}(k+1), \\ \qquad\qquad \text{with probability } 1 - \frac{1}{N}\sum_{i=1}^m a_i(k) + \mathcal{O}(1/N^2) \\ \sigma\mathbf{X}(k) + \mathbf{R}(k+1) - \mathbf{e}_i, \text{ with probability } \frac{1}{N}a_i(k) + \mathcal{O}(1/N^2) \\ \text{“other”, with probability } \mathcal{O}(1/N^2) \end{cases}$$

where

$$a_i(k) \equiv X_{i+1}(k)R_i(k+1) + \binom{R_i(k+1)}{2}. \tag{1}$$

Here σ is the shift operator defined by

$$\sigma(X_1(k), \ldots, X_m(k)) \equiv (X_2(k), \ldots, X_m(k), 0)$$

and $\mathbf{e}_i$ is the ith unit vector.

Remark. The above model has some similarities to island models and their structured coalescents (cf. Wilkinson-Herbots 1998 and Notohara 1990). If we think of having m islands, the urn model is equivalent to the following. Start with n individuals. Each generation, the individuals in islands $2, \ldots, m$ move lock-step one island to the left (preserving their “slots” on the islands).

No coalescing happens during this shifting. The individuals who were in island 1 are relocated to the m islands according to the above probabilities. If the slot they land on is occupied, the two individuals coalesce. Thus we can think of a migration model in which the simple shifting migrants cannot take part in coalescing; only the "island 1" migrants are allowed to coalesce.

Put

$$\Delta A_N(k) = A_N(k+1) - A_N(k) = -\#(\text{coalescence events in step } k).$$

It follows that

$$\begin{aligned}\mathbb{P}(\Delta A_N(k) = -1 | \mathbf{X}(k), \mathbf{R}(k+1)) &= \frac{1}{N}\sum_{i=1}^{m-1} X_{i+1}(k) R_i(k+1) \\ &\quad + \frac{1}{N}\sum_{i=1}^{m} \binom{R_i(k+1)}{2} + \mathcal{O}(\frac{1}{N^2}).\end{aligned}$$

Hence, computing the expected values of $\mathbf{R}(k+1)$ given $\mathbf{X}(k)$,

$$\begin{aligned}\mathbb{P}(\Delta A_N(k) = -1 | \mathbf{X}(k)) &= \frac{1}{N} X_1(k) \sum_{i=1}^{m-1} X_{i+1}(k)\, b_i \\ &\quad + \frac{1}{N}\binom{X_1(k)}{2} \sum_{i=1}^{m} b_i^2 + \mathcal{O}(\frac{1}{N^2}). \qquad (2)\end{aligned}$$

Here, when averaging over the values of $\mathbf{R}(k+1)$, we have used the fact that, given $\mathbf{X}(k)$, $R_i(k+1) \sim \text{Bin}(X_1(k), b_i)$. Using (2), we see that the coalescence probability in state $(x_1, \ldots, x_m) \in S_r$ is

$$\begin{aligned}&\mathbb{P}(\text{coalescent event at next step} \,|\, \mathbf{X} = (x_1, \ldots, x_m)) \\ &\quad = \frac{1}{N}\left\{ x_1 \sum_{i=1}^{m-1} x_{i+1}\, b_i + \binom{x_1}{2} \sum_{i=1}^{m} b_i^2 \right\} + \mathcal{O}(\frac{1}{N^2}). \qquad (3)\end{aligned}$$

Next we want to bring in the stationary distribution for the configuration process and justify its appearance in the $N \to \infty$ limit. It turns out that the time between coalescent events will be long enough so that, for large N, the configuration process will reach an equilibrium on each level before the coalescence occurs. These equilibrium distributions on the different levels are studied next.

For a given $r \in \{1, \ldots, n\}$, if we consider the configuration process $\mathbf{X}(k)$ conditioned to be in level r and to experience no coalescences, we get the **level-r configuration process**, $\mathbf{X}^{(r)}(k)$, $k = 1, 2, \ldots$. This process develops according to

$$X_j^{(r)}(k+1) = X_{j+1}^{(r)}(k) + R_j^{(r)}(k+1) \qquad (j = 1, \ldots, m), \qquad (4)$$

where $X_{m+1}^{(r)}(k) = 0$ and, conditionally on $X_1^{(r)}(k)$,

$$\mathbf{R}^{(r)}(k+1) = (R_1^{(r)}(k+1), \ldots, R_m^{(r)}(k+1)) \sim \text{Mult}(X_1^{(r)}(k); b_1, \ldots, b_m).$$

To get the stationary distribution for the level-r configuration process, we introduce the probabilities

$$\beta_j = \frac{\mathbb{P}(B \geq j)}{E(B)} = \frac{\sum_{i=j}^m b_i}{\sum_{i=1}^m i b_i} \qquad (j = 1, \ldots, m).$$

It is clear that $\beta_1 + \cdots + \beta_m = 1$, that $\beta_j = \beta_1(b_j + \cdots + b_m)$, and that the β_j's satisfy the system of equations

$$\begin{aligned} \beta_j &= \beta_{j+1} + \beta_1 b_j \quad (j = 1, \ldots, m-1) && (5) \\ \beta_m &= \beta_1 b_m. && (6) \end{aligned}$$

An interpretation, borrowed from renewal theory, is that the current age of a seed picked from a randomly chosen generation would follow the probabilities β_j, thus forming the seed bank equilibrium age distribution.

Lemma 1 *The unique stationary distribution for the level-r configuration process is given, for each $r \in \{1, \ldots, n\}$, by*

$$\mathbf{X}^{(r)}(\infty) = (X_1^{(r)}(\infty), \ldots, X_m^{(r)}(\infty)) \sim \text{Mult}(r; \beta_1, \ldots, \beta_m).$$

Proof. Suppose $\mathbf{X}^{(r)}(k) \sim \text{Mult}(r; \beta_1, \ldots, \beta_m)$; equivalently it has probability generating function

$$\mathbb{E}\left[u_1^{X_1^{(r)}(k)} \cdots u_m^{X_m^{(r)}(k)}\right] = (u_1\beta_1 + \cdots + u_m\beta_m)^r.$$

Using the recursion (4), we get

$$\mathbb{E}\left[u_1^{X_1^{(r)}(k+1)} \cdots u_m^{X_m^{(r)}(k+1)}\right] = \mathbb{E}\left[u_1^{X_2^{(r)}(k)} \cdots u_{m-1}^{X_m^{(r)}(k)}\, u_1^{R_1^{(r)}(k+1)} \cdots u_m^{R_m^{(r)}(k+1)}\right].$$

Next, we condition on $X_1^{(r)}(k)$, use the relation

$$\mathbb{E}\left[u_1^{R_1^{(r)}(k+1)} \cdots u_m^{R_m^{(r)}(k+1)} | X_1^{(r)}(k)\right] = (b_1u_1 + \cdots + b_mu_m)^{X_1^{(r)}(k)},$$

and the fact that $(X_2^{(r)}(k), \ldots, X_m^{(r)}(k))$ and $(R_1^{(r)}(k+1), \ldots, R_m^{(r)}(k+1))$ are conditionally independent, given $X_1^{(r)}(k)$. This yields

$$\begin{aligned} &\mathbb{E}\left[u_1^{X_1^{(r)}(k+1)} \cdots u_m^{X_m^{(r)}(k+1)}\right] \\ &= \mathbb{E}\left\{\mathbb{E}\left[u_1^{X_2^{(r)}(k)} \cdots u_{m-1}^{X_m^{(r)}(k)} | X_1^{(r)}(k)\right] \mathbb{E}\left[u_1^{R_1^{(r)}(k+1)} \cdots u_m^{R_m^{(r)}(k+1)} | X_1^{(r)}(k)\right]\right\} \\ &= \mathbb{E}\left[(b_1u_1 + \cdots + b_mu_m)^{X_1^{(r)}(k)} u_1^{X_2^{(r)}(k)} \cdots u_{m-1}^{X_m^{(r)}(k)}\right] \\ &= [\beta_1(b_1u_1 + \cdots + b_mu_m) + \beta_2u_1 + \cdots + \beta_mu_{m-1}]^r \\ &= (\beta_1u_1 + \cdots + \beta_mu_m)^r, \end{aligned}$$

the last line following from the equations (5) and (6) for the β_j's. Thus the multinomial distribution $\text{Mult}(r; \beta_1, \ldots, \beta_m)$ is a stationary distribution for $\mathbf{X}^{(r)}(\cdot)$.

Finally, the fact that this Markov chain is irreducible implies that the stationary distribution is unique. ∎

Up to now, everything has been rigorous. We finish our derivation of the theorem with the following heuristic ideas which, while intuitively clear, need to be justified (cf. next section). First note that transitions occur within level r (due to "migration") during each step, whereas coalescence events occur with probability $\mathcal{O}(1/N)$ in a given step (i.e., it takes $\mathcal{O}(N)$ steps to get a coalescence). Taking N large, this gives the configuration process time to reach equilibrium between coalescence events. This should allow us to use the stationary distribution of the configuration process on each level. In the seedbank model, the proportion of time in any state of a given level is given by the stationary distribution at that state, independent of the starting state, because the process reaches stationarity long before a coalescence occurs. Lemma 1 tells us that the (long-run) proportion of time spent in state $(x_1, \ldots, x_m) \in S_r$ by the configuration process (while it is on level r) is

$$\frac{r!}{x_1! \cdots x_m!} \beta_1^{x_1} \cdots \beta_m^{x_m}.$$

Using standard results on multinomial random variables, if

$$\mathbf{X} = (X_1, \ldots, X_m) \sim \text{Mult}(r; \beta_1, \ldots, \beta_m),$$

then

$$\mathbb{E}(X_1 X_{i+1}) = 2\binom{r}{2}\beta_1\beta_{i+1} \qquad (i = 1, \ldots, m-1)$$

and

$$\mathbb{E}\binom{X_1}{2} = \binom{r}{2}\beta_1^2.$$

Now the coalescence probability in (3) suggests that, on the coalescent time scale (speed up time by a factor of N), the *coalescence rate* while the configuration process is in state $(x_1, \ldots, x_m)$ should be

$$\rho(x_1, \ldots, x_m) \equiv x_1 \sum_{i=1}^{m-1} x_{i+1}\, b_i + \binom{x_1}{2} \sum_{i=1}^{m} b_i^2. \tag{7}$$

Weighting these rates by the proportion of time in each state of level r, we see that the coalescence rate when there are r ancestors should be

$$\begin{aligned}
\mathbb{E}(\rho(\mathbf{X})) &= \mathbb{E}\left[X_1 \sum_{i=1}^{m-1} b_i X_{i+1} + \binom{X_1}{2} \sum_{i=1}^{m} b_i^2\right] \\
&= \sum_{i=1}^{m-1} 2\beta_1\beta_{i+1}\binom{r}{2} b_i + \beta_1^2 \binom{r}{2} \sum_{i=1}^{m} b_i^2 \\
&= \beta_1^2 \binom{r}{2} \left(2 \sum_{i=1}^{m-1} \frac{\beta_{i+1}}{\beta_1} b_i + \sum_{i=1}^{m} b_i^2\right) \\
&= \beta_1^2 \binom{r}{2}, \qquad (8)
\end{aligned}$$

as

$$\begin{aligned} 2\sum_{i=1}^{m-1}\frac{\beta_{i+1}}{\beta_1}b_i+\sum_{i=1}^{m}b_i^2 &= 2\sum_{i=1}^{m-1}(b_{i+1}+\cdots+b_m)b_i+\sum_{i=1}^{m}b_i^2 \\ &= 2\sum_{1\le i<j\le m}b_ib_j+\sum_{i=1}^{m}b_i^2 \\ &= (b_1+\cdots+b_m)^2=1. \end{aligned}$$

Thus, the seedbank coalescent should be the usual coalescent run on a slower time scale. Note that if $m = 1$ (i.e., no seed bank), we would have $\beta_1 = 1$ and hence this reduces to the usual coalescence rate.

3 Proof of the Theorem.

We now justify the heuristic part of the derivation in the previous section and, in addition, include the ingredients necessary for a rigorous proof of weak convergence. Since we want convergence to a relatively simple type of process (continuous-time pure death process), weak convergence will follow from Theorem 2.12 on p. 173 of Ethier and Kurtz (1986) if we can show convergence of 1-dimensional distributions. In their theorem we let

1. $E = \{1, ..., n\}$ be the "ancestor space,"
2. $E_N = S$ for all finite N, where $S = S_1 \cup \cdots \cup S_n$ is configuration space,
3. $\eta_N : S \to E$ defined by $\eta_N(x) = |x| = x_1 + \cdots + x_m$ for any configuration $x = (x_1, ..., x_m)$. (I.e., η_N is just the projection which maps a configuration onto the number of ancestors in that configuration.)
4. $\mathbf{X}_N(k), k = 0, 1, ...$ is the discrete-time configuration process (when population size is N),
5. $A_N([Nt]) = \eta_N(\mathbf{X}_N([Nt])) = |\mathbf{X}_N([Nt])|$ is the number of ancestors in the speeded up configuration process.

Order the states in S so that level 1 states occur first,..., level n states occur last, the ordering of states within a level being arbitrary but fixed. With this ordering, let $\mathbf{\Pi}_N = (\mathbf{\Pi}_N(x, y))_{x,y\in S}$ be the 1-step transition probability matrix for the configuration process $\mathbf{X}_N(k)$ when the population size is N. Finally, set $T(t)f(i) = \mathbb{E}^i f(A(t))$, where $A(t)$ is the coalescent process described in Theorem 1, and the superscript i on the expectation refers to the initial state.

According to the theorem of Ethier and Kurtz, to show weak convergence of $A_N([N\cdot])$ to $A(\cdot)$, it is enough to show that, for each function $f : E \to \mathbb{R}$ and each configuration x,

$$\left|\sum_{y\in S}\mathbf{\Pi}_N^{[Nt]}(x, y)f(|y|) - T(t)f(|x|)\right| \to 0, \text{ as } N \to \infty.$$

Here, $\mathbf{\Pi}_N^k(x,y)$ is the (x,y) term in the kth power $\mathbf{\Pi}_N^k$ of the transition matrix for the configuration process.

In the above limit, for any $i \in \{1,\ldots,n\}$,

$$T(t)f(i) = \sum_{j=1}^{n} f(j) \sum_{k=0}^{\infty} \frac{t^k}{k!} q_{ij}^{(k)},$$

where $Q = (q_{ij})$ is the $n \times n$ generator matrix for the desired coalescent (cf. Theorem 1) and $q_{ij}^{(k)}$ is the (i,j) element of Q^k. Also

$$\sum_{y \in S} \mathbf{\Pi}_N^{[Nt]}(x,y) f(|y|) = \sum_{j=1}^{n} f(j) \sum_{y \in S_j} \mathbf{\Pi}_N^{[Nt]}(x,y).$$

The function f in both equations maps $\{1,...,n\} \to \mathbb{R}$. Since the first sum involves only a finite number of terms, it is enough to prove for each $i \in \{1,...,n\}$ and $x \in S_i$, and each $j \in \{1,...,n\}$ that

$$\left| \sum_{y \in S_j} \mathbf{\Pi}_N^{[Nt]}(x,y) - \sum_{k=0}^{\infty} \frac{t^k}{k!} q_{ij}^{(k)} \right| \to 0, \text{ as } N \to \infty. \tag{9}$$

The second sum is just the (i,j) element of the matrix e^{tQ}.

Remark. To assist the reader in the matrix calculations, we will use the following notational convention. When a matrix C corresponds to transitions in configuration space S, we write $C(x,y)$ for the (x,y) element (with the aforementioned ordering of states). When a matrix D corresponds to transitions in the "collapsed space" $E = \{1,...,n\}$, we write $d_{i,j}$ for the (i,j) element. Finally, the notation $C_{i,j}(x,y)$ will represent the (x,y) element of the sub-matrix $C_{i,j}$ in a larger block matrix. As a general rule, we will use boldface letters to denote the large matrices corresponding to the full set of states in S; the submatrices making up these larger matrices will be in regular type.

The proof of (9) will be based on a result of MÖHLE (1998) which will help us handle the asymptotics of $\mathbf{\Pi}_N^{[Nt]}$ as $N \to \infty$. We begin by using the calculations of the previous section to write

$$\mathbf{\Pi}_N = \mathbf{A} + \frac{1}{N}\mathbf{B} + \mathcal{O}(\frac{1}{N^2}) \tag{10}$$

where

$$\mathbf{A} = \begin{bmatrix} A_{11} & 0 & 0 & \cdots & 0 & 0 \\ 0 & A_{22} & 0 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ & & & \ddots & & \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & A_{n-1,n-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & A_{n,n} \end{bmatrix},$$

and

$$\mathbf{B} = \begin{bmatrix} -B_{11} & 0 & 0 & \cdots & 0 & 0 & 0 \\ B_{21} & -B_{22} & 0 & \cdots & 0 & 0 & 0 \\ 0 & . & . & \cdots & . & . & . \\ & & & \ddots & & & \\ . & . & . & & . & . & . \\ 0 & 0 & 0 & \cdots & B_{n-1,n-2} & -B_{n-1,n-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & B_{n,n-1} & -B_{n,n} \end{bmatrix}.$$

Here, $A_{i,i}$ is the 1-step transition matrix for the level-i configuration process (i.e., ignoring coalescence events) and the $B_{i,j}$'s are matrices that come about due to single coalescence events.

Since the size of level k is

$$d_k \equiv |S_k| = \binom{k+m-1}{m-1},$$

$A_{k,k}$ must be a $d_k \times d_k$ matrix, $B_{k,k}$ must be a $d_k \times d_k$ matrix, and $B_{k,k-1}$ must be a $d_k \times d_{k-1}$ matrix. The overall matrix $\mathbf{\Pi}_N$ is $d \times d$, where

$$d \equiv |S| = \sum_{k=1}^{n} \binom{k+m-1}{m-1} = \binom{m+n}{m} - 1$$

is the size of the configuration space. In the above matrices, each 0 denotes a zero matrix (of the appropriate size).

Now Lemma 1 of Möhle (1998) implies

$$\lim_{N\to\infty} \mathbf{\Pi}_N^{[Nt]} = \lim_{N\to\infty} \left(\mathbf{A} + \frac{1}{N}\mathbf{B} + \mathcal{O}(\frac{1}{N^2}) \right)^{[Nt]} = \mathbf{P} - \mathbf{I} + e^{t\mathbf{G}}, \qquad (11)$$

where $\mathbf{P}$ is the projection matrix $\mathbf{P} \equiv \lim_{k\to\infty} \mathbf{A}^k$, $\mathbf{I}$ is the $d \times d$ identity matrix, and $\mathbf{G} \equiv \mathbf{PBP}$.

We know from Lemma 1 that the rows of the resulting block matrices $P_{i,i} = \lim_{k\to\infty} A_{i,i}^k$ are all the same and, moreover, that these rows are given by the probabilities for a Mult$(i; \beta_1, \ldots, \beta_m)$ distribution. More specifically, if $y^{(1)}, y^{(2)}, \ldots, y^{(d_i)}$ are the ordered elements of S_i, then each row of $P_{i,i}$ is given by the vector

$$\mathbf{M}_i \equiv \left(M_i(y^{(1)}), M_i(y^{(2)}), \ldots, M_i(y^{(d_i)}) \right) \qquad (12)$$

where, for $y = (y_1, y_2, \ldots, y_m) \in S_i$, we define the multinomial probability

$$M_i(y) \equiv \frac{i!}{y_1! \cdots y_m!} \beta_1^{y_1} \cdots \beta_m^{y_m}.$$

Note that, in this notation, (8) becomes

$$\sum_{y\in S_i} M_i(y)\rho(y) = \beta_1^2 \binom{i}{2}. \qquad (13)$$

If $x \in S_i$, then the corresponding row in the matrix $\mathbf{B}$ has non-zero elements only within the sub-matrix $B_{i,i-1}$, whose columns correspond to configurations

$y \in S_{i-1}$, or within $B_{i,i}$, whose columns correspond to configurations $y \in S_i$. By listing all possible coalescence events starting from a configuration $x \in S_i$, it follows that each rowsum in $B_{i,i-1}$ adds up to the quantity $\rho(x)$ defined in (7):

$$\sum_{y \in S_{i-1}} B_{i,i-1}(x,y) = \rho(x). \tag{14}$$

The nonzero columns of $B_{i,i}$ and $B_{i,i-1}$ consist of the same vectors, even though they are not typically in the same locations within these matrices. Hence

$$\sum_{y \in S_i} B_{i,i}(x,y) = \rho(x). \tag{15}$$

If these steps are not clear, the reader is urged to work through a simple example (say with $m = 2$ and $n = 2$) to see how the matrices break down.

Note that

$$\mathbf{G} = \begin{bmatrix} -G_{11} & 0 & 0 & \cdots & 0 & 0 \\ G_{21} & -G_{22} & 0 & \cdots & 0 & 0 \\ 0 & . & . & \cdots & . & . \\ & & \ddots & & & \\ . & . & . & . & . & . \\ 0 & 0 & \cdots & G_{n-1,n-2} & -G_{n-1,n-1} & 0 \\ 0 & 0 & \cdots & 0 & G_{n,n-1} & -G_{n,n} \end{bmatrix},$$

where

$$\begin{aligned} G_{i,i} &= P_{i,i} B_{i,i} P_{i,i} \\ G_{i,i-1} &= P_{i,i} B_{i,i-1} P_{i-1,i-1}. \end{aligned}$$

These matrices simplify because of the particular structure of the factors.

Since all rows in $P_{i,i}$ are equal and given by $\mathbf{M}_i$ in (12), we obtain for $x, y \in S_i$,

$$\begin{aligned} G_{i,i}(x,y) &= \sum_{u \in S_i} \sum_{v \in S_i} P_{i,i}(x,u) B_{i,i}(u,v) P_{i,i}(v,y) \\ &= M_i(y) \sum_{u \in S_i} P_{i,i}(x,u) \rho(u) \\ &= \beta_1^2 \binom{i}{2} M_i(y), \end{aligned}$$

where we have used (15) and (13). Since $P_{i,i}(x,y) = M_i(y)$ for all $x \in S_i$, we conclude that

$$G_{i,i} = c_i \, P_{i,i}, \qquad \text{where } c_i \equiv \beta_1^2 \binom{i}{2}. \tag{16}$$

Similarly,

$$G_{i,i-1} = c_i \, \tilde{P}_{i,i-1}, \tag{17}$$

where $\tilde{P}_{i,i-1}$ is the $d_i \times d_{i-1}$ matrix with identical rows given by the probability vector $\mathbf{M}_{i-1}$ appearing in $P_{i-1,i-1}$.

Notation. Write $(\mathbf{G}^k)_{i,j}$ for the (i,j) block in the matrix $\mathbf{G}^k$ and $G_{i,j}^k$ for the matrix $G_{i,j}$ raised to the power k.

The matrix powers $\mathbf{G}^k$ are lower triangular block matrices given recursively by

$$(\mathbf{G}^k)_{i,i} = (-1)^k G_{i,i}^k = (-1)^k c_i^k P_{i,i}^k = (-1)^k c_i^k P_{i,i}\,, \tag{18}$$

$$(\mathbf{G}^k)_{i,j} = G_{i,i-1}(\mathbf{G}^{k-1})_{i-1,j} - G_{i,i}(\mathbf{G}^{k-1})_{i,j}\,, \quad j = 1, \dots i-1, \tag{19}$$

and, of course, $(\mathbf{G}^k)_{i,j} = 0$ when $j > i$. Note that we have used the fact that $P_{i,i}$ is a projection and hence $P_{i,i}^k = P_{i,i}$.

We need to calculate

$$\lim_{N\to\infty} \sum_{y\in S_j} \mathbf{\Pi}_N^{[Nt]}(x,y) = \sum_{y\in S_j} (\mathbf{P} - \mathbf{I} + e^{t\mathbf{G}})(x,y) = \sum_{y\in S_j} (e^{t\mathbf{G}})(x,y)$$

and show that it is equal to $(e^{tQ})_{ij}$ for all $x \in S_i$. By definition,

$$\sum_{y\in S_j} (e^{t\mathbf{G}})(x,y) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \sum_{y\in S_j} \mathbf{G}^k(x,y).$$

Furthermore, for any $x \in S_i$,

$$\sum_{y\in S_j} \mathbf{G}^k(x,y) = \sum_{y\in S_j} (\mathbf{G}^k)_{i,j}(x,y),$$

where the quantity on the right hand side does not depend on the particular x but only on $|x| = i$.

For $j = i$, by (16) and (18),

$$\sum_{y\in S_i} (\mathbf{G}^k)_{i,i}(x,y) = (-1)^k c_i^k \sum_{y\in S_i} P_{i,i}(x,y) = (-1)^k c_i^k. \tag{20}$$

For $j = 1, \dots, i-1$, by (17) and (19),

$$\begin{aligned}
\sum_{y\in S_j} (\mathbf{G}^k)_{i,j}(x,y) &= \sum_{y\in S_j} (G_{i,i-1}(\mathbf{G}^{k-1})_{i-1,j} - G_{i,i}(\mathbf{G}^{k-1})_{i,j})(x,y) \\
&= c_i \sum_{y\in S_j} \left(\tilde{P}_{i,i-1}(\mathbf{G}^{k-1})_{i-1,j} - P_{i,i}(\mathbf{G}^{k-1})_{i,j}\right)(x,y) \\
&= c_i \sum_{y\in S_j} (\mathbf{G}^{k-1})_{i-1,j}(x,y) - c_i \sum_{y\in S_j} (\mathbf{G}^{k-1})_{i,j}(x,y).
\end{aligned} \tag{21}$$

The last line follows from the fact that any given block $(\mathbf{G}^{k-1})_{i,j}$ of $\mathbf{G}^{k-1}$ will have identical rows. For example, it is easy to check that

$$(\mathbf{G}^2)_{i,i} = c_i^2 P_{i,i},$$

$$(\mathbf{G}^2)_{i,i-1} = -c_i^2 P_{i,i}\tilde{P}_{i,i-1} - c_i c_{i-1}\tilde{P}_{i,i-1}P_{i-1,i-1},$$

and

$$(\mathbf{G}^2)_{i,i-2} = c_i c_{i-1}\tilde{P}_{i,i-1}\tilde{P}_{i-1,i-2},$$

and clearly each of these matrices consists of identical rows. Note also that, to avoid notationally cumbersome resizing of the matrices in the first sum of the

right hand side of (21), we take $\sum_{y\in S_j}(\mathbf{G}^{k-1})_{i-1,j}(x,y)$ to mean the common rowsum in the $d_{i-1}\times d_j$ matrix $(\mathbf{G}^{k-1})_{i-1,j}$, which is the same for any choice of x in S_i.

Equations (20) and (21) show that, for each k, the quantities

$$f_k(i,j)\equiv\sum_{y\in S_j}\mathbf{G}^k(x,y),\quad x\in S_i$$

are well defined and satisfy the recursive system of equations

$$\begin{aligned} f_k(i,i) &= (-1)^k c_i^k, \qquad k\geq 1\\ f_k(i,j) &= c_i f_{k-1}(i-1,j)-c_i f_{k-1}(i,j),\quad i\geq 2,\ 1\leq j\leq i-1,\quad k\geq 2. \end{aligned}$$

This is the same linear system of equations as that satisfied by the elements $q_{i,j}^{(k)}$ of the matrix Q^k. By uniqueness of the solution of this system, we may therefore make the identification

$$f_k(i,j)=q_{i,j}^{(k)},$$

and hence

$$\sum_{y\in S_j}(e^{t\mathbf{G}})(x,y)=(e^{tQ})_{|x|,j}.$$ ∎

4 Adding Mutations.

We now study the effects of neutral mutations in the seed bank model. We will keep the derivation on a more intuitive level, leaving the task of enlarging the state space even more (to accomodate the methods of the previous section) to the industrious and indefatigable reader. In general, the mutation probability $u_N(j)$ for an allele produced by a seed which is j generations old will depend on the age of the seed. Typically, older seeds are more likely to have mutations, so we have

$$u_N(1)\leq u_N(2)\leq\cdots\leq u_N(m),$$

where, as before, N is the population size and m is the number of generations seeds can remain viable. Since our model is neutral, we will not keep track of the type of mutation that occurs. It is enough to record when a mutation occurs. The type of mutation can be read off from a transition probability matrix, depending on the application.

We take the usual diffusion scaling

$$u_N(j)=\frac{\theta(j)}{2N}\quad(j=1,\ldots,m).$$

With this scaling, the probability of multiple mutations in a given generation vanishes as $N\to\infty$. Thus, on the coalescent time scale, $\theta(j)/2$ gives the mutation rate for individuals produced by age j seeds.

A given individual will have arisen from an age j seed (i.e., corresponds to a relocation to cell j) with probability b_j. We mark such an individual with a

mutation at the time of the relocation with probability $u_N(j)$. Thus the overall mutation probability for an ancestor being "relocated" is

$$b_1 u_N(1) + b_2 u_N(2) + \cdots + b_m u_N(m).$$

It should be clear that mutations only occur during relocations since these correspond to birth events forward in time. Thus the mutation probability (for the whole set of ancestors) depends on the current configuration; in particular, on the number $X_1(k)$ of individuals in cell 1 at any time k. We know that β_1 is the stationary probability that a given ancestral line is currently in cell 1. Thus the overall mutation probability for a given ancestral line is

$$\beta_1 \left(b_1 u_N(1) + b_2 u_N(2) + \cdots + b_m u_N(m)\right)$$

per generation. Multiplying by N (i.e., measuring time in units of N generations), we get the *mutation rate*

$$\gamma \equiv \frac{\beta_1}{2} \left(b_1\theta(1) + b_2\theta(2) + \cdots + b_m\theta(m)\right) = \frac{\beta_1}{2}\overline{\theta}$$

along a given ancestral line in the coalescent, where $\overline{\theta} \equiv b_1\theta(1) + b_2\theta(2) + \cdots + b_m\theta(m)$. Such mutations are independent along different branches in the coalescent tree.

This suggests that we will see a different pattern of variability in seed bank models. In particular, the time for two ancestors to coalesce in our model is $T_2 \sim \text{Exp}(\beta_1^2)$, so the probability that they will be identical by descent is

$$\begin{aligned} \mathbb{P}(IBD) &= \mathbb{E}[\mathbb{P}(IBD|T_2)] \\ &= \int_0^\infty \beta_1^2 e^{-\beta_1^2 t} e^{-2\gamma t} dt \\ &= \frac{1}{1 + 2\gamma\beta_1^{-2}} = \frac{1}{1 + \overline{\theta}\beta_1^{-1}}, \end{aligned} \tag{22}$$

where γ is the above mutation rate. This should be compared to the classical formula

$$\mathbb{P}(IBD) = \frac{1}{1+\theta}$$

when there is no seed bank and $\theta = \theta(1)$ is the only relevant mutation rate.

Note that if all $\theta(i) \equiv \theta$ (i.e., no age-dependence for seed mutation rate), then $\overline{\theta} = \theta$, so for a nontrivial seed bank model ($\beta_1 < 1$),

$$\mathbb{P}(IBD) = \frac{1}{1 + \theta\beta_1^{-1}} < \frac{1}{1+\theta}.$$

So, in this case, the presence of the seed bank increases variability. There will be even more variability, of course, when the mutation probability increases with seed age.

Acknowledgements: IK was supported in part by NFR grant M9481-650. SK was supported in part by NSF grant DMS-00-72198. Some of the work on this paper was done while ML was visiting the Department of Zoology, Hong Kong University. He would like to thank Dr. Mei Sun for her hospitality during his stay. Finally, we express our thanks to the referee for some useful suggestions.

References

[1] Alvarez-Buylla, E.R., Chaos, A., Pinero, D. and Garay, A.A. (1996). Demographic genetics of a pioneer tropical tree species: path dynamics, seed dispersal, and seed banks. *Evolution* **50**: 1155-1166.

[2] Donnelly, P. (1999). The coalescent and microsatellite variability. In: *Microsatellites: Evolution and applications*. Goldstein, D.B. and Schlötterer, C. (eds.) Oxford University Press pp. 116-128.

[3] Donnelly, P. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics* **29**: 401-421.

[4] Epling, C., Lewis, H., and Ball, F.M. (1960). The breeding group and seed storage: a study in population dynamics. *Evolution* **14**: 238-255.

[5] Ethier, S.N. and Kurtz, T.G. (1986) *Markov Processes: Characterization and Convergence*. Wiley, New York.

[6] Gomez, A. and Carvalho G.R. (2000). Sex, parthenogenesis and genetic structure of rotifers: microsatellite analysis of contemporary and resting egg bank populations. *Mol. Ecol.* **9**: 203-214.

[7] Hairston Jr., N.G. and De Stasio Jr., B.T. (1988). Rate of evolution slowed by a dormant propagule pool. *Nature* **336**: 239-242.

[8] Kingman, J.F.C. (1982). The coalescent. *Stoch. Proc. Appl.* **13**: 235-248.

[9] Levin, D.A. (1990). The seed bank as a source of genetic novelty in plants. *Am. Nat.* **135**: 563-572.

[10] McCue, K.A. and Holtsford, T.P. (1998). Seed bank influences on genetic diversity in the rare annual Clarkia springvillensis (Onagraceae). *Am. J. Bot.* **85**: 30-36.

[11] Möhle, M. (1998). A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Adv. Appl. Prob.* **30**: 493-512.

[12] Möhle, M. (2000). Ancestral processes in population genetics – the coalescent. *J. Theor. Biol.* **204**: 629-638.

[13] Nordborg, M. (2001). Coalescent theory. To appear in *Handbook of Statistical Genetics*. Balding, D.J., Cannings, C. and Bishop, M. (eds.) Wiley, Chichester.

[14] Notohara, M. (1990). The coalescent and the genealogical process in geographically structured populations. *J. Math. Biol.* **29**: 59-75.

[15] Templeton, A.R. and Levin, D.A. (1979). Evolutionary consequences of seed pools. *Am. Nat.* **114**: 232-249.

[16] Tonsor, S.J., Kalisz, S., Fisher, J., and Holstford, T.P. (1993). A life-history based study of population genetic structure: seed bank to adults in Plantago lanceolata. *Evolution* **47**: 833-843.

[17] Wilkinson-Herbots, H.M. (1998). Genealogy and subpopulation differentiation under various models of population structure. *J. Math. Biol.* **37**: 535-585.

Figure 1: The urn model. (a) The $m-$window at step 0. The $m-$window at step 0 includes cell 0 to cell $m-1$. (b) Sliding the window and relocating the balls from the previous leftmost cell, solid balls meaning that two balls have coalesced. (c) The resulting $m-$window at step 1.