Theoretical Population Biology 74 (2008) 104–114

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms

Ori Sargsyan *, John Wakeley

Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, 02138, USA

ARTICLE INFO

Article history:
Received 8 March 2008
Available online 14 May 2008

Keywords:
Coalescent
Simultaneous multiple mergers
Marine organisms

ABSTRACT

We describe a forward-time haploid reproduction model with a constant population size that includes life history characteristics common to many marine organisms. We develop coalescent approximations for sample gene genealogies under this model and use these to predict patterns of genetic variation. Depending on the behavior of the underlying parameters of the model, the approximations are coalescent processes with simultaneous multiple mergers or Kingman's coalescent. Using simulations, we apply our model to data from the Pacific oyster and show that our model predicts the observed data very well. We also show that a fact which holds for Kingman's coalescent and also for general coalescent trees – that the most-frequent allele at a biallelic locus is likely to be the ancestral allele – is not true for our model. Our work suggests that the power to detect a "sweepstakes effect" in a sample of DNA sequences from marine organisms depends on the sample size.

1. Introduction

Coalescent processes provide a framework for studying genetic variation in a random sample of DNA sequences. These gene-genealogical models describe the impact of the genetic drift on the ancestry of a sample. Genetic variation is regularly observed in samples from most species, and interest in explaining this variation focuses attention on the ancestry of the sample. However, the most natural starting point for population genetic analyses is to consider evolutionary processes in the entire population, forward in time. Backward-time, genealogical processes are derived from forward-time reproduction models. The Wright-Fisher model and the Moran model are two commonly used forward-time reproduction models, while the Cannings models provide a more general setting; see Cannings (1974, 1975). Kingman (1982a,b) showed that for a subset of the Cannings models, and for the Moran model, the genealogy of the sample can be approximated by a coalescent process when the variance of the number of offspring of an individual, σ_N^2, has a finite limit as $N \to \infty$. We adopt the common terminology and call this Kingman's coalescent.

Life history characteristics common to many marine organisms can lead to a high variance of offspring number, because few individuals may contribute to highly-successful reproduction events. Many marine organisms, from a wide variety of taxonomic groups, have the potential for great fecundity but also suffer high early mortality (Hedgecock, 1994). Breeding adults release gametes into the water column where fertilization takes place. The larvae are planktonic. Successful reproduction is a chance event, depending on oceanographic conditions favorable to spawning, fertilization, larval survival and successful recruitment. The majority of larvae are lost due to predation and other factors such as currents carrying them away from suitable benthic habitats. Disturbance can also be an important factor in opening up habitat patches for re-colonization (Witman, 1987; Dayton, 1971; Paine and Levin, 1981). Another feature of these organisms that bears upon modeling efforts is that generations overlap. Adults live for a number of years, and successful breeding events may come only in the most favorable years. These life history characteristics put constraints on the family sizes, such that a few individuals may have very large family sizes while others will have a few offspring or none (Hedgecock, 1994; Beckenbach, 1994).

Kingman's coalescent is a powerful tool for understanding patterns of genetic variation in a sample of DNA sequences. However, it may not be appropriate for marine organisms that have the above life history characteristics, because they may violate the underlying assumptions of the model, in particular regarding σ_N^2. There is some evidence for this in genetic data. For example, the haplotype-frequency distribution in a sample of mitochondrial DNA (mtDNA) data from Pacific oyster shows an excess of common haplotypes, a deficiency of haplotypes at intermediate frequencies and too many singleton haplotypes (Beckenbach, 1994; Boom et al., 1994; Reeb and Avise, 1990). As pointed out by Beckenbach (1994), these patterns are different from the patterns that are predicted from Ewens' sampling theory (Ewens, 1972), which holds when the genealogy of a sample can be approximated by Kingman's

* Corresponding address: Harvard University, 4096 Biological Laboratories, 16 Divinity Ave., Cambridge, MA 02138, USA.
E-mail address: sargsyan@fas.harvard.edu (O. Sargsyan).

0040-5809/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2008.04.009

coalescent. Thus, it is of interest to develop other coalescent approximations for cases in which σ_N^2 is very large, meaning not bounded as $N \to \infty$.

To address these concerns, Beckenbach (1994), Flowers et al. (2002), Wakeley and Takahashi (2003), Hedrick (2005), and Eldon and Wakeley (2006), have all considered forward-time reproduction models and their predictions about genetic variation in a sample. With the exception of Wakeley and Takahashi (2003) and Eldon and Wakeley (2006), however, all of these studies were restricted to forward-time dynamics. Eldon and Wakeley (2006) took a coalescent approach and derived all possible coalescent approximations for the genealogy of a sample under one particular model. Here, we take a similar approach: first we define a simple, forward-time reproduction model which includes the above life history characteristics, then we study it using coalescent methods. We relax the condition (Eldon and Wakeley, 2006) that only a single individual can have a large family at any one time. We also do not assume particular functional forms for the parameters in the limit $N \to \infty$.

The problem of approximating the genealogy of a sample under Cannings' models, for cases in which σ_N^2 is not bounded, has been studied by several authors. In the limit of large population size N, different ancestral processes are obtained depending how the moments of the distribution of the number of offspring per individual behave as N grows. While the Kingman coalescent includes only binary mergers of ancestral lines, other models result in processes with multiple mergers that are separated in time (Pitman, 1999; Sagitov, 1999), or with simultaneous multiple mergers (Möhle and Sagitov, 2001; Sagitov, 2003; Schweinsberg, 2000). We use these results for classifying the coalescent approximations under our model.

After developing the coalescent framework for the model in the next section, we compare predictions from our model to those of Kingman's coalescent. Analytical results concerning patterns of genetic variation are difficult to obtain from multiple-mergers coalescent processes, so many of our results were obtained using the simulation algorithm we develop below. First, we show that our model predicts well the patterns of the mtDNA sequence variation in Pacific oysters described above. Second, we use simulations to suggest that, under models such as ours, the most frequent allele is probably *not* the ancestral allele, in stark contrast to the predictions of Kingman's coalescent. Third, we show that the ability to detect the signature of a sweepstake effect (Hedgecock, 1994) in a sample of DNA sequences from marine organisms, or from other species with large variance of offspring numbers, may depend strongly on the sample size. Very large sample sizes may be needed to reliably distinguish multiple-mergers coalescent processes from the simple Kingman coalescent.

2. Methods and results

We begin with a forward-time reproduction model for a finite population that includes life history characteristics common to many marine organisms. The population size N is constant over time, there are no selective effects, and time unfolds in discrete steps. At time t, $(X_N, Y_N)_t$ is a random variable with joint distribution $\pi_N(\cdot,\cdot)$, X_N and Y_N being integers between 1 and N, the number of individuals who will not survive and who take place in the reproductive process respectively. For simplicity, we assume that the joint distribution $\pi_N(\cdot,\cdot)$ does not change over time. We select the X_N individuals and the Y_N individuals independently and at random from the population. X_N new individuals are generated at random as per a Wright-Fisher model: each selects as its parent one of the Y_N potential parents, and these replace the X_N of the population chosen to die. When $X_N = 1$ and $Y_N = 1$ this is the Moran model. We shall refer to the case where the model does not have $X_N = Y_N = 1$ as an SRE (Successful Reproduction Event) and we suppose the Moran Model applies with probability $1 - \epsilon_N$, and the SRE with probability ϵ_N.

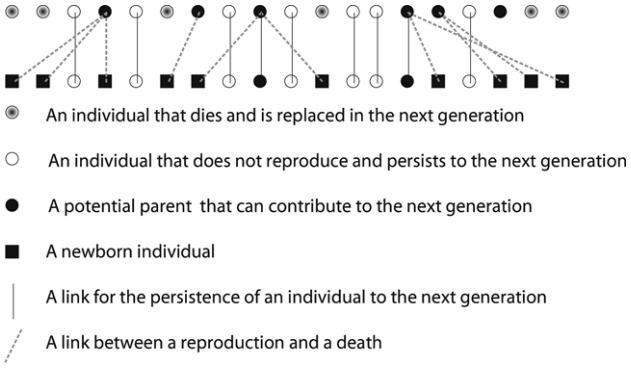

Fig. 1. An example of a successful reproduction event (a single step in the forward-time process). Here, the population size N is 16, the number of potential parents Y_N is 6, and the number of new born individuals X_N is 10.

The biological interpretation of our model is as follows. Broadcast spawning occurs at regular intervals. Each individual produces a very large number of offspring, enough to potentially replace a large fraction of the population. Normally, with probability $1 - \epsilon_N$, there is little space for these newborns: one adult dies and is replaced by a single offspring. Occasionally, with probability ϵ_N, a disturbance event occurs that kills X_N adult individuals, allowing X_N offspring to settle and survive (an SRE). Depending on the current conditions, not every adult has an equal chance to be the parent of these X_N offspring. We assume that Y_N adults are the only potential parents of the X_N offspring. Not every disturbance event is the same: The numbers X_N and Y_N are drawn from a joint distribution $\pi_N(\cdot,\cdot)$ and the particular individuals chosen to die and to reproduce are selected uniformly at random from the population. This leads to several types of individuals which are shown in Fig. 1.

Note that in the above model the family sizes are exchangeable due to the random choice of the Y_N and X_N individuals. In contrast to the models in Cannings (1974), in our model generations are overlapping. This distinction will become important when we consider mutation, as we will only allow newborns to mutate. The Wright-Fisher and Moran models are special cases of the above model when $Y_N = X_N = N$, $\epsilon_N = 1$; and when $Y_N = 1$ and $X_N = 1$ (see above) or $\epsilon_N = 0$, respectively.

2.1. Ancestral processes

We apply the results of Möhle and Sagitov (2001) and Sagitov (2003) for characterizing the limiting ancestral processes for a sample under our model. Note that their results are for Cannings' models with non-overlapping generations but in our model we have overlapping generations. However, we can still apply their results here because our model can be converted into a Cannings model (see Cannings (1973)), with limiting ancestral processes identical to those we describe below, simply by replacing the persistence events in our model with reproduction-and-death events (see Fig. 1). The exception to this is in Section 2.2 when we consider mutation, which we will allow to occur only in newborn individuals.

The results of Möhle and Sagitov (2001) and Sagitov (2003) are summarized in Table 1 in the form of the following limits: $\mathbb{P}_1(1,1) \to C_0$, $\frac{\mathbb{P}_1(1,1,1)}{\mathbb{P}_1(1,1)} \to C_1$ and $\frac{\mathbb{P}_2(2,2)}{\mathbb{P}_1(1,1)} \to C_2$ as $N \to \infty$. The quantity $\mathbb{P}_1(1,1)$ is the probability that two randomly chosen individuals coalesce in one time step; $\mathbb{P}_1(1,1,1)$ is the probability that three randomly chosen individuals coalesce in one time

Table 1
The limit conditions for approximating the genealogies of samples by the coalescent processes

Conditions	Approximation of the ancestral process
$C_0 = 0$	A continuous coalescent process
$C_0 \neq 0$	A discrete Markov chain process
$C_0 = 0$ and $C_1 = 0$	Kingman's coalescent
$C_0 = 0$, $C_1 \neq 0$ and $C_2 = 0$	Coalescent with simultaneous mergers
$C_0 = 0$, $C_1 \neq 0$ and $C_2 \neq 0$	Coalescent with multiple mergers

step; and $\mathbb{P}_2(2,2)$ is the probability that four randomly chosen individuals have two parents and each of these two parents has two descendants in these four, again in one time step. We derive conditions on the limits $\frac{X_N}{N} \Longrightarrow \phi$, $Y_N \Longrightarrow Y$, and $\epsilon_N \longrightarrow \epsilon$, by which we can classifying the various limiting ancestral processes. By "$\Longrightarrow$" we mean convergence in distribution and by "$\rightarrow$" we mean uniform convergence (here simply of the real number ϵ_N).

Because we will consider two possibilities for Y_N: $Y_N \Longrightarrow Y$ and $Y_N \rightarrow \infty$, we adopt the notion $\mathbb{P}(Y = \infty) = 0$ for $Y_N \Longrightarrow Y$. This is to emphasize that Y is an integer valued random variable. We do not consider the cases when the definition of the limit in probability, $Y_N \Longrightarrow Y$, includes the case of discrete random variables Y that have positive mass at ∞. An example of such a case is when $\mathbb{P}(Y_N = 2) = \mathbb{P}(Y_N = N/2) = 0.5$ then Y, such that,

$$\mathbb{P}(Y = 2) = \mathbb{P}(Y = \infty) = 0.5,$$

can be considered as the limit of Y_N in probability. Such cases may fall under the second possibility above, that $Y_N \rightarrow \infty$. Also, we use the notion $\mathbb{P}(\phi = 0) = 0$ for $\frac{X_N}{N} \Longrightarrow \phi$ to emphasize that ϕ does not have point mass at zero and to separate from the case when $\frac{X_N}{N} \rightarrow 0$.

For our model, we obtain

$$\begin{aligned}\mathbb{P}_1(1,1) &= \epsilon_N \mathbb{E}\left(\frac{X_N(X_N-1)}{N(N-1)}\frac{1}{Y_N} + 2\frac{X_N}{N}\frac{N-X_N}{N-1}\frac{1}{N}\right) \\ &\quad + (1-\epsilon_N)\frac{2}{N}\frac{1}{N-1}, \end{aligned} \tag{1}$$

$$\begin{aligned}\mathbb{P}_1(1,1,1) &= \epsilon_N \mathbb{E}\left(\frac{X_N(X_N-1)(X_N-2)}{N(N-1)(N-2)}\frac{1}{Y_N^2}\right. \\ &\quad \left. + 3\frac{X_N}{N}\frac{X_N-1}{N-1}\frac{N-X_N}{N-2}\frac{1}{NY_N}\right), \end{aligned} \tag{2}$$

and

$$\begin{aligned}\mathbb{P}_2(2,2) = \epsilon_N \mathbb{E}&\left(\frac{3X_N(X_N-1)(X_N-2)(X_N-3)}{N(N-1)(N-2)(N-3)}\frac{(Y_N-1)}{Y_N^3}\right. \\ &+ \frac{12X_N(X_N-1)(X_N-2)(N-X_N)}{N(N-1)(N-2)(N-3)}\frac{(Y_N-1)}{NY_N^2} \\ &\left. + \frac{12X_N(X_N-1)(N-X_N)(N-X_N-1)}{N(N-1)(N-2)(N-3)}\frac{(Y_N-1)}{N(N-1)Y_N}\right). \end{aligned} \tag{3}$$

From the above equations we have the following approximations when N is very large.

$$\begin{aligned}\mathbb{P}_1(1,1) &\approx \epsilon_N \mathbb{E}\left(\frac{X_N^2}{N^2}\left(\frac{X_N-1}{X_N Y_N} + \frac{2(1-\frac{X_N}{N})}{X_N}\right)\right) \\ &\quad + (1-\epsilon_N)\frac{2}{N^2}, \end{aligned} \tag{4}$$

$$\mathbb{P}_1(1,1,1) \approx \epsilon_N \mathbb{E}\left(\frac{X_N^2(X_N-1)}{N^3 Y_N}\left(\frac{X_N-2}{X_N Y_N} + \frac{3(1-\frac{X_N}{N})}{X_N}\right)\right), \tag{5}$$

and

$$\begin{aligned}\mathbb{P}_2(2,2) \approx \epsilon_N \mathbb{E}&\left(\frac{3X_N(X_N-1)(X_N-2)(X_N-3)}{N^4}\frac{(Y_N-1)}{Y_N^3}\right. \\ &+ \frac{12X_N(X_N-1)(X_N-2)(N-X_N)}{N^4}\frac{(Y_N-1)}{NY_N^2} \\ &\left. + \frac{12X_N(X_N-1)(N-X_N)(N-X_N-1)}{N^4}\frac{(Y_N-1)}{NY_N}\right). \end{aligned} \tag{6}$$

From here on, all limits and approximations should be understood to hold as "$N \rightarrow \infty$" or "for large N" (i.e. we will usually skip these phrases).

Since our model is a mixture of two reproduction processes, the probability of coalescence per time step, $\mathbb{P}_1(1,1)$, is the sum of two probabilities: that the two individuals coalesce in the background of an SRE or that the two individuals coalesce in the background of a Moran-model reproduction event.

The first case we consider is when a coalescent event is much more likely to happen in the background of an SRE than in a Moran-model reproduction event. In particular, we assume that $\mathbb{P}(Y = \infty) = 0$, $\mathbb{P}(\phi = 0) = 0$ and $\epsilon_N \rightarrow \epsilon$, $\epsilon \neq 0$. From Eq. (4) it follows that under these conditions, when N is large coalescent events occur almost exclusively in SREs. Then, from Eqs. (4) to (6) we have the limits

$$\mathbb{P}_1(1,1) \rightarrow \mathbb{E}\left(\frac{\epsilon\phi^2}{Y}\right),$$

$$\mathbb{P}_1(1,1,1) \rightarrow \mathbb{E}\left(\frac{\epsilon\phi^3}{Y^2}\right),$$

and

$$\mathbb{P}_2(2,2) \rightarrow \mathbb{E}\left(\frac{\epsilon\phi^4(Y-1)}{Y^3}\right).$$

From these results, and according to Table 1, it follows that the ancestry of a sample can be approximated by a discrete Markov chain with simultaneous multiple mergers; for more details see Sagitov (2003). Note that when Y is a constant and $\phi = 1$ and $\epsilon_N = 1$ we have the model considered by Wakeley and Takahashi (2003).

The second case we consider (which will be the basis for the simulation algorithms presented in Section 2.3) is when a coalescent event may happen in either background. That is, we assume that $\mathbb{P}(Y = \infty) = 0$, $\mathbb{P}(\phi = 0) = 0$, $\epsilon_N \rightarrow 0$, and $N^2\epsilon_N \rightarrow c$, where $c \neq 0$. We allow c to be ∞, which is equivalent to saying $N^2\epsilon_N \rightarrow \infty$. From Eq. (4) we have

$$\mathbb{P}_1(1,1) \approx \epsilon_N \mathbb{E}\left(\frac{\phi^2}{Y} + \frac{2}{c}\right). \tag{7}$$

In other words, the relative rates of coalescence in the background of an SRE and in the background of Moran-model reproduction event are $\mathbb{E}\left(\frac{\phi^2}{Y}\right)$ and $2/c$, respectively. Note that when $c = \infty$ then coalescence will happen only in background of an SRE. From Eqs. (5) to (6) we have the following limits for the other probabilities:

$$\frac{\mathbb{P}_1(1,1,1)}{\mathbb{P}_1(1,1)} \rightarrow \frac{\mathbb{E}\left(\frac{\phi^3}{Y^2}\right)}{\mathbb{E}\left(\frac{\phi^2}{Y} + \frac{2}{c}\right)}, \tag{8}$$

and

$$\frac{\mathbb{P}_2(2,2)}{\mathbb{P}_1(1,1)} \rightarrow \frac{\mathbb{E}\left(\frac{\epsilon\phi^4(Y-1)}{Y^3}\right)}{\mathbb{E}\left(\frac{\phi^2}{Y} + \frac{2}{c}\right)}. \tag{9}$$

According to Table 1, it follows that the ancestry of a sample can be approximated by a continuous-time coalescent process with simultaneous multiple mergers. Note that when $Y \neq 1$ there will be simultaneous mergers, while if $Y = 1$ there will only be multiple mergers, asynchronous in time. The case $Y = 1, \phi \neq 0$ and $\epsilon_N = 1/N^\gamma$ was studied by Eldon and Wakeley (2006).

Another possibility for the underlying parameters of the model is the case when coalescent events are much more likely to happen in the background of Moran-model reproduction events. Here, we assume that $\mathbb{P}(Y = \infty) = 0$, $\mathbb{P}(\phi = 0) = 0$ and $N^2\epsilon_N \to 0$. In this case, from (4) we have $\mathbb{P}_1(1, 1) \approx \frac{2}{N^2}$. Therefore, the ratio $\mathbb{P}_1(1, 1, 1)/\mathbb{P}_1(1, 1)$ converges to zero. Note that for this case the condition $N^2\epsilon_N \to 0$ simply means that all coalescent events for a finite sample will occur in the background of the Moran process, before the first SRE. This occurs because SREs occur on average every $1/\epsilon_N$ time steps but a coalescent event in the background of the Moran process takes on average $N^2/2$ steps. In this case, our model behaves like the Moran model. Therefore, the ancestral history of a sample can be approximated by Kingman's coalescent.

Finally, we consider the possibility that coalescent events may occur in either background but, despite the fact that SREs may dominate the ancestry of a sample, nonetheless the ancestral process becomes Kingman's coalescent. In particular, we assume that $Y_N \to \infty$ or $X_N/N \to 0$. Note that we can use uniform convergence for random variables because they can be treated as functions and it is stronger than convergence in distribution. We show that the ancestral history of a sample can be approximated by Kingman's coalescent; a proof is given in Appendix A. Note that this means that the time scale of the resulting Kingman coalescent may be very different that the usual time scale, which is proportional to N generations.

2.2. Mutation processes

We assume that only newborn individuals can mutate. Individuals who simply persist to the next time step do not mutate. Each newborn experiences a mutation with probability μ, and we further assume that mutations occur according to the infinitely many sites model, in which every mutation occurs at a unique site. To model the mutation process for a sample we need to trace the ancestry of a single individual. It follows directly from our model that the times between consecutive mutational events on one lineage looking backwards in time are geometrically distributed.

Our goal in this section is to show how the mutation process should be scaled and what should be assumed about mutation probability, such that the mutation process can be approximated by a Poisson process. Sometimes this involves making the well-justified assumption that the mutation probability per nucleotide site per birth event is small, while the details of the modeling depend on the relative sizes of μ, N, and other parameters, in the limit $N \to \infty$. Interestingly, a Poisson mutation process can be obtained in some case even without making the usual assumption that $\mu \to 0$. We express all results for the mutation rate per lineage as $\theta/2 = \text{const} \times \theta_0/2$, where $\theta_0/2$ captures our assumptions about μ and the limiting constant depends on the other parameters of the model.

In all of the following, we restrict ourselves to cases in which the genealogy of the sample can be approximated by a continuous-time coalescent process. That is, we assume that $\mathbb{P}_1(1, 1) \to 0$ holds. When we trace the mutational history of a single individual back into the past, the number of time steps back to the first mutation event on the lineage will be distributed as a geometric random variable with the following non-success probability per step:

$$\epsilon_N \mathbb{E}\left(\frac{N - X_N}{N} + \frac{X_N}{N}(1-\mu)\right) + (1-\epsilon_N)\left(\frac{N-1}{N} + (1-\mu)\frac{1}{N}\right)$$

or, after simplification,

$$1 - \mathbb{E}\left(\epsilon_N \frac{X_N}{N} + (1-\epsilon_N)\frac{1}{N}\right)\mu.$$

Since we scale the ancestral process of a sample size n by the inverse of the coalescence probability, $1/\mathbb{P}_1(1, 1)$, we will scale the mutational process (backwards in time) by this same factor. To approximate the mutation process by a Poisson process after this scaling by $1/\mathbb{P}_1(1, 1)$, we require that

$$\mathbb{E}\left(\epsilon_N \frac{X_N}{N} + (1-\epsilon_N)\frac{1}{N}\right)\mu/\mathbb{P}_1(1, 1) \tag{10}$$

has a finite, non-zero limit. In the limit, Eq. (10) becomes the time-rescaled rate of mutation per lineage. Following the usual notation, we denote this limit by $\theta/2$. We derive conditions for μ, depending on the behavior of X_N, Y_N, and ϵ_N, to achieve a finite limit for (10).

The first case we consider is when simultaneous multiple mergers occur in the limiting ancestral process. Recall that in this case, we assume that

$$\mathbb{P}(Y = \infty) = 0, \qquad \mathbb{P}(\phi = 0) = 0, \quad \epsilon_N \to 0 \text{ and } c > 0,$$

where $N^2\epsilon_N \to c$; and c could be ∞. Under these conditions on the parameters from Eq. (7) we have $\mathbb{P}_1(1, 1) \approx \epsilon_N \mathbb{E}\left(\frac{\phi^2}{Y} + \frac{2}{c}\right)$ and

$$\frac{\left(\epsilon_N \mathbb{E}\frac{X_N}{N} + (1-\epsilon_N)\frac{1}{N}\right)\mu}{\mathbb{P}_1(1, 1)} \approx \frac{\mu\left(\epsilon_N \mathbb{E}\phi + (1-\epsilon_N)\frac{1}{N}\right)}{\epsilon_N \mathbb{E}\left(\frac{\phi^2}{Y} + \frac{2}{c}\right)}. \tag{11}$$

It is helpful here to define another quantity: a is equal to the limit of $N\epsilon_N$ as $N \to \infty$. Depending on the limits $N^2\epsilon_N \to c$ and $N\epsilon_N \to a$ we have the following possibilities:

(1) $0 < c < \infty$ (or $c = \infty$ ($N^2\epsilon_N \to \infty$) and $a = 0$ ($N\epsilon_N \to 0$)). For this case it follows that $a = 0$ therefore from Eq. (11), the limit in (10) is finite if $\mu\frac{1}{N\epsilon_N}$ has a finite limit. For the case $0 < c < \infty$ the limit requirement becomes that μN has finite limit $\theta_0/2$ and

$$\frac{\theta}{2} = \frac{\theta_0}{2}\frac{1}{c\mathbb{E}\left(\frac{\phi^2}{Y} + \frac{2}{c}\right)}.$$

If $N^2\epsilon_N \to \infty$ and $N\epsilon_N \to 0$ then $\theta_0/2$ is the limit of $\mu\frac{1}{N\epsilon_N}$ and

$$\frac{\theta}{2} = \frac{\theta_0}{2}\frac{1}{\mathbb{E}\left(\frac{\phi^2}{Y}\right)}.$$

Note that for the above case we need to assume that $\mu \to 0$ holds.

(2) $0 < a < \infty$. Under this condition $N^2\epsilon_N \to \infty$ and from Eq. (11), it follows that we do not need to assume that μ goes to zero, and we have

$$\frac{\theta}{2} = \frac{\mu\left(\mathbb{E}\phi + \frac{1}{a}\right)}{\mathbb{E}\left(\frac{\phi^2}{Y}\right)}.$$

The second case we consider is when the genealogy of a sample can be approximated by Kingman's coalescent. Let us assume that $X_N \to X$, $Y_N \to \infty$ or $Y_N \Rightarrow Y$, where $\mathbb{P}(Y = \infty) = 0$ and $\epsilon_N \to \epsilon$. Under these conditions, from Eq. (4), we have

$$\mathbb{P}_1(1, 1) \approx \frac{1}{N^2}\left(\epsilon\mathbb{E}\left(\frac{X(X-1)}{Y} + 2X\right) + 2(1-\epsilon)\right)$$

and

$$\left(\epsilon_N \mathbb{E}\frac{X_N}{N} + (1-\epsilon_N)\frac{1}{N}\right)\mu \approx \frac{\mu}{N}(\epsilon\mathbb{E}(X-1) + 1).$$

Therefore, in this case the finite-limit condition in (10) is equivalent to $N\mu$ having a finite limit $\theta_0/2$. In the limit we have

$$\frac{\theta}{2} = \frac{\theta_0}{2}\frac{(\epsilon\mathbb{E}(X-1)+1)}{(\epsilon\mathbb{E}\,(X(X-1)/Y+2X)+2(1-\epsilon))},$$

where we assume that $\mathbb{E}X^2 < \infty$.

A third case for the parameters is $\epsilon_N = \frac{1}{N}$, $Y_N \to \infty$, $\frac{Y_N}{N} \to y$, $\mathbb{P}(\phi = 0) = 0$ and Y_N is a constant. If $y = 0$, then using approximation in (1) under these conditions we have

$$\mathbb{P}_1(1,1) \approx \frac{\mathbb{E}\phi^2}{NY_N}$$

and

$$\left(\epsilon_N\mathbb{E}\frac{X_N}{N} + (1-\epsilon_N)\frac{1}{N}\right)\mu \approx \frac{\mathbb{E}\phi+1}{N}\mu.$$

Thus the finite limit condition in (10) is equivalent to $Y_N\mu$ having a finite limit $\theta_0/2$, and

$$\frac{\theta}{2} = \frac{\theta_0}{2}\frac{(\mathbb{E}\phi+1)}{\mathbb{E}\phi^2}.$$

If $y > 0$, then we have

$$\mathbb{P}_1(1,1) \approx \frac{1}{N^2}\mathbb{E}\left(\phi^2/y + 2\phi(1-\phi)+2\right)$$

and

$$\left(\epsilon_N\mathbb{E}\frac{X_N}{N} + (1-\epsilon_N)\frac{1}{N}\right)\mu \approx \frac{\mathbb{E}\phi+1}{N}\mu,$$

and hence

$$\frac{\theta}{2} = \frac{\theta_0}{2}\frac{(\mathbb{E}\phi+1)}{\mathbb{E}\left(\phi^2/y + 2\phi(1-\phi)+2\right)},$$

where $\theta_0/2$ is the limit of μN.

2.3. Simulation algorithms

Here, we describe two simulation algorithms for the cases when the genealogy of a sample can be approximated by a continuous-time coalescent process (that is, we assume $\mathbb{P}_1(1,1) \to 0$) with simultaneous multiple mergers (or simply multiple mergers). Both algorithms generate the genealogy of a sample without mutations. The mutation process is treated separately, as a Poisson process with rate $\theta/2$ along each branch of the gene genealogy.

In the second case in Section 2.1, we obtained a continuous coalescent process with simultaneous multiple mergers when $\epsilon_N \to 0$ and $N^2\epsilon_N \to c$, $0 < c \le \infty$, $\mathbb{P}(\phi = 0) = 0$, and $\mathbb{P}(Y = \infty) = 0$. In this case $\mathbb{P}_1(1,1)$ is proportional to ϵ_N, following Eq. (7). Time is measured in units of $1/\epsilon_N$ (rather than $1/\mathbb{P}_1(1,1)$) time steps in the simulations. With this time scale, we can approximate the occurrence of SREs by a Poisson process with rate one. Note that after simulating the genealogy of the sample, the $1/\mathbb{P}_1(1,1)$ time scale can be retrieved by multiplying the coalescence times by $\mathbb{E}\left(\frac{\phi^2}{Y} + \frac{2}{c}\right)$.

Our first algorithm is for $N^2\epsilon_N \to \infty$ ($c = \infty$) in which case the ancestry is dominated by SREs.

Algorithm 1. (1) Start with $K = n$. (K is the number of the ancestors of the sample as we follow them back in time.)
(2) Sample (ϕ, Y) from a joint distribution $\pi(\cdot,\cdot)$.
(3) Generate the waiting time t to the first SRE (looking backwards in time): It is an exponential random variable with mean one.
(4) Sample $(k_0, k_1, \ldots, k_Y)$ from $Mn(K, p_0, p_1, \ldots, p_Y)$, where $\sum_{i=0}^{Y} k_i = K$, $p_0 = 1-\phi$ and $p_i = \phi/Y$, $i = 1, \ldots, Y$. (Here $Mn(\cdot)$ denotes the multinomial distribution.)
(5) If $k_i \neq 0, i = 1, \ldots, Y$ then randomly choose k_i individuals from K and merge them into a common ancestor.
(6) Reset the number of the ancestors to $K = k_0 + \sum_{i=1}^{Y} \mathbf{1}\{k_i \neq 0\}$, where $\mathbf{1}\{x\}$ is one if the x is true and zero, otherwise. Stop if K is equal to one, otherwise return to Step 2.

Our second algorithm is for $0 < c < \infty$, in which case coalescent events can occur via either an SRE or a Moran-model reproduction event.

Algorithm 2. (1) Start with $K = n$.
(2) Sample (ϕ, Y) from a joint distribution $\pi(\cdot,\cdot)$.
(3) Generate the waiting time t to the first SRE (looking backwards in time): It is an exponential random variable with mean one.
(4) Up to time t, run the usual binary, Kingman coalescent with rate $K(K-1)/c$ while there are K lineages. Form pairwise common ancestors as coalescent events occur, and decrement K by one at each event. Exit the algorithm if $K = 1$ before time t.
(5) Using the value of K (> 1) from Step 4, sample $(k_0, k_1, \ldots, k_Y)$ from $Mn(K, p_0, p_1, \ldots, p_Y)$, where $p_0 = 1-\phi$ and $p_i = \phi/Y$, $i = 1, \ldots, Y$.
(6) If $k_i \neq 0, i = 1, \ldots, Y$ then randomly choose k_i individuals from K without replacement and coalesce them together.
(7) Reset K to $K = k_0 + \sum_{i=1}^{Y} \mathbf{1}\{k_i \neq 0\}$, where $\mathbf{1}\{x\}$ is one if the x is true and zero, otherwise. Stop if K is equal to one, otherwise go to Step 2.

2.4. Simulation results

Using the simulation algorithms described above, we provide three examples of how predictions about genetic variation in a sample can be different depending on whether the genealogy is given by Kingman's coalescent or by the coalescent with simultaneous multiple mergers derived from our model. We implemented the above algorithms in a program written in the C programming language, which is available from the authors upon request.

Example 1. We consider the data from Boom et al. (1994). The data consist of restriction-site variation in a sample of mitochondrial DNA sequences from Pacific oyster (*Crassostrea gigas*) from British Columbia. Using 9 restriction enzymes, 44 haplotypes were discovered among 141 samples. For our analysis we assume the gain or loss of a restriction site is a unique event. Adopting this assumption in an analysis of the fragment sizes (see Table 2 in Boom et al. (1994)) we find that there are 48 segregating sites in the sample. We summarize the data in two different ways. First, we count the frequency of each haplotype, and rank order these. Second, we classify each segregating site by the number of times the minor allele at that site is observed in the sample; these counts are often called the site-frequency spectrum.

We apply a simplified version of our model to the data, assuming that (ϕ, Y) are constants. As in the simulation algorithms above, we consider the case where $N^2\epsilon_N \to c$, $0 < c < \infty$ and $N\epsilon_N \to 0$. We use a likelihood approach to estimate the underlying parameters (ϕ, Y, θ, c) of the model, first assuming that the site-frequency counts represent counts of mutant alleles (unfolded case below), that is assuming that the most-frequent allele is the ancestral allele at each site. However, as we see in Example 2, this assumption is not necessarily true for populations that have been evolving according our model. Thus, we also analyze the data as folded site frequencies (folded case below), that is assuming we do not know the ancestral type.

We use Algorithm 2 to compute the likelihood of the data for a particular set of values of the parameters, (ϕ, Y, θ, c). For the

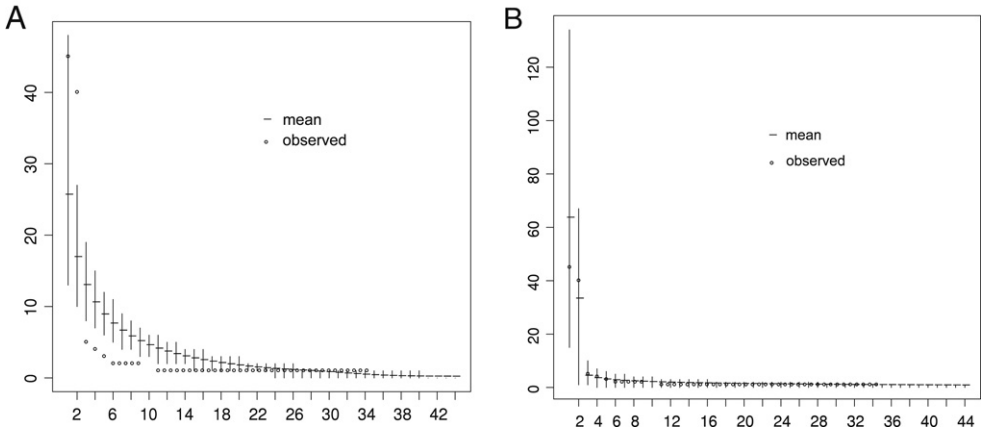

Fig. 2. The upper and lower end points of vertical segments represent 2.5% and 97.5% quantiles of the marginal distributions of rank ordered haplotype frequencies. We used "−" for the means of marginal frequencies and "o" for observed frequencies. (A) Haplotype frequency spectrum under Kingman's coalescent versus observed data. Mutation rate is $\theta_K = 12$. (B) Haplotype-frequency spectrum under simultaneous multiple mergers. The values of the parameters are the following: $\phi = 1, Y = 2, \theta = 1.55$, $c = 125$. The sample size is 141.

unfolded case, first we generate a gene genealogy of the sample of size $n = 141$, then we compute the sums of the lengths of the branches in the tree that have $i, i = 1, \ldots, n-1$, descendants in the sample. We denote these L_i. For the folded case, we compute sums of the lengths of the branches that have i or $n-i$ descendants. When $2i \neq n$ the lengths are $L_i + L_{n-i}$, for $i = 1, \ldots, \lfloor n/2 \rfloor$, where $\lfloor x \rfloor$ is the integer part of x; but in the special case $2i = n$ the lengths are simply L_i.

Given the gene genealogy, the mutation process is an independent Poisson process along the branches of the tree. The rate of mutation is $\theta/2$ along each branch. The probability that there are n_i segregating sites with frequency i is

$$\mathbb{P}_i^{(1)} = \frac{(\theta L_i/2)^{n_i}}{n_i!} e^{-\theta L_i/2}.$$

For the folded case, the probability, that the number of sites with frequencies i or $n-i$ is $n_i + n_{n-i}$, is given by the following expressions: if $2i \neq n$, then

$$\mathbb{P}_i^{(2)} = \frac{(\theta(L_i + L_{n-i})/2)^{n_i + n_{n-i}}}{(n_i + n_{n-i})!} e^{-\theta(L_i + L_{n-i})/2};$$

if $2i = n$, then

$$\mathbb{P}_i^{(2)} = \frac{(\theta L_i/2)^{n_i}}{n_i!} e^{-\theta L_i/2}.$$

Note that we make the usual assumption above, that $0! = 1$. For convenience, if both n_i and L_i are equal to zero we define each of the above probabilities to be equal to one. The probability of the data $(n_1, \ldots, n_{n-1})$ given the tree is the product $\mathbb{P}_1 \cdots \mathbb{P}_{n-1}$, since given the tree the mutation process is independent along the edges of the tree. In the folded case, the probability is the product of $\mathbb{P}_i^{(2)}$, $i = 1, \ldots, \lfloor n/2 \rfloor$. We estimate the likelihood by taking the average of the products of these probabilities over a large number of gene genealogies generated with Algorithm 2.

We generated 100 000 gene genealogies for every possible combination of the following parameter values:

$\phi = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$,

$Y = \{1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13\}$,

$c = \{.1, 2, 5, 10, 50, 100, 125, 150, 175, 200, 225, 250, 300, 400, 500, 1000, 10\,000\}$,

$\theta_1 = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.5, 2, 2.5, 3, 3.5, 4\}$,

where $\theta_1 = \theta_0/c \approx 2\mu N/c$.

The mutation rate θ under Algorithm 2 is given by the formula

$$\frac{\theta}{2} = \frac{\theta_1}{2} \frac{1}{\mathbb{E}\left(\frac{\phi^2}{Y} + \frac{2}{c}\right)}.$$

The maximum likelihood parameter estimates were identical in both the unfolded and folded cases: $\phi = 1, Y = 2, c = 125, \theta_1 = 0.8$, and $\theta = 1.55$. Expected number of segregating sites and the expected number of haplotypes are 39.31 and 31.41, respectively, for estimated values of the parameters for our model.

To get a sense of what this means for the genealogy of the sample, when $\phi = 1$, $Y = 2$, $c = 125$, SREs occur with rate equal to one, while binary mergers due to Moran-model reproduction events occur $c/2 = 62.5$ times more slowly than in the Kingman coalescent. Therefore, many ancestral lineages will likely be present when an SRE occurs. When an SRE occurs and $\phi = 1$ the entire population is replaced by the offspring of the Y successful parents. With $Y = 2$, the k ancestral lineages that are present will merge into the two ancestors, each with equal probability. The number of descendant-lineages of one parent will follow the Binomial$(1/2, k)$ distribution, and the rest of the lineages will be descended from the other parent.

We also applied the above approach under Kingman's coalescent, to estimate its single parameter, the mutation rate θ. The estimates for unfolded and folded cases were 12 and 8, respectively. The expected number of segregating sites and the expected number of haplotypes are 66.36 and 30.98 for $\theta = 12$; 44.18 and 23.84 for $\theta = 8$, respectively.

To further quantify the predictions of our model and of the Kingman coalescent, we computed the haplotype-frequency spectrum and the site-frequency spectrum for each of 100 000 samples of size 141 using the parameter estimates above. The haplotype-frequency spectrum is defined as the rank-ordered frequencies of the haplotypes in a sample. To assess the variability of our predictions, we used the 100 000 pseudo-samples to estimate the 2.5% and 97.5% quantiles for each frequency class. We say that the model predicts the observed data well if the observed haplotype-frequency counts and site-frequency counts fall within the (2.5%, 97.5%) quantile-intervals obtained from the simulations.

The comparisons for the haplotype-frequency spectrum are represented in Fig. 2. Under Kingman's coalescent, some of the haplotype frequencies are outside the (2.5%, 97.5%) quantile-intervals (panel A). In particular, under Kingman's coalescent it appears unlikely to observe two haplotypes with frequencies as high as those of the two most-frequent haplotypes in the sample. Under our model, however, the observed rank ordered haplotype

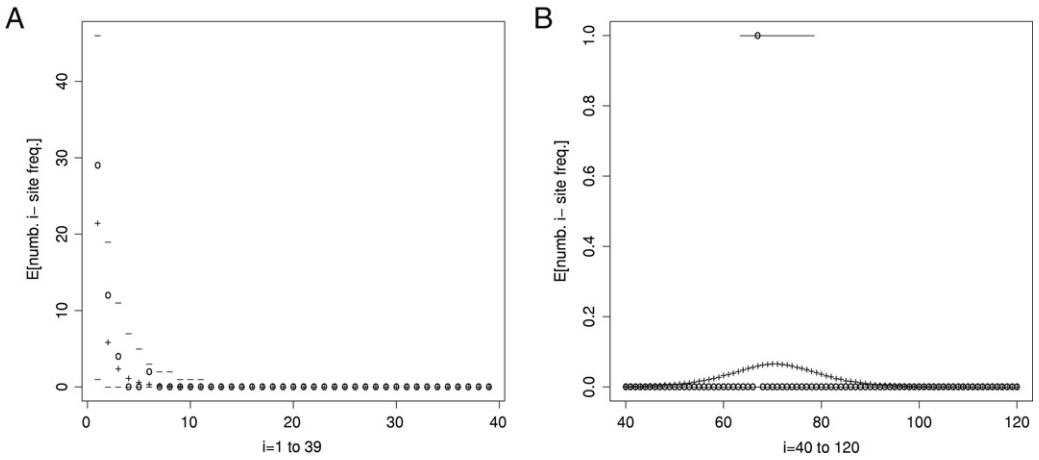

Fig. 3. Site frequency spectrum of a sample of size 141 under the simultaneous multiple mergers model. The values of the parameters are $\phi = 1, Y = 2, \theta = 1.55, \theta_1 = 0.8$, and $c = 125$. Points marked respectively: "o" for observed data; "−" for 2.5% and 97.5% quantiles and "+" for the means of marginal frequencies. (A) the site-frequency spectrum for frequencies from 2 to 39; (B) site-frequencies from 40 to 120.

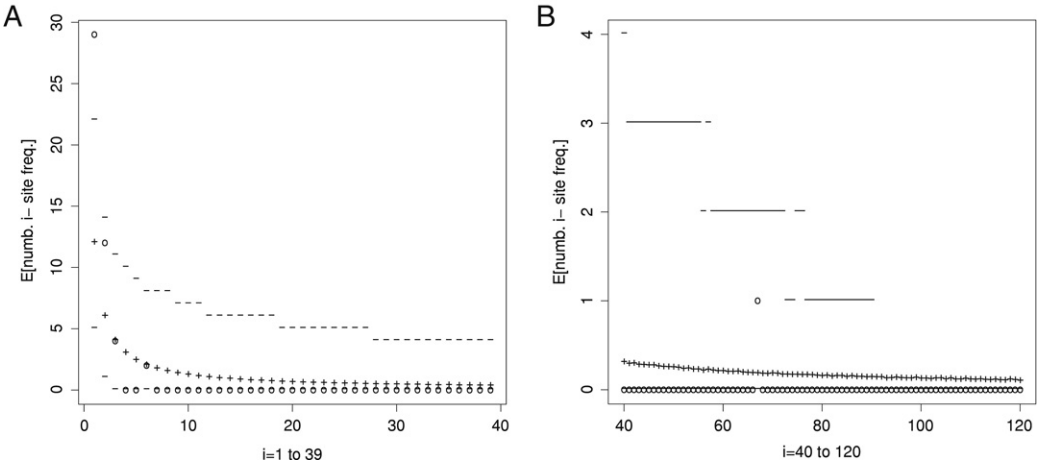

Fig. 4. Site frequency spectrum for a sample size 141 under Kingman's coalescent. The mutation rate is $\theta = 12$. Points marked "o" for observed data; "−" for 2.5% and 97.5% quantiles and "+" for the means of the marginal frequencies. (A) the site-frequency spectrum for frequencies from 2 to 39; (B) site-frequencies from 40 to 120.

frequencies are all in the (2.5%, 97.5%) quantile-intervals (panel B). In our model, with $Y = 2$, it is not unlikely to observe two frequent haplotypes.

A similar type of result occurs for the site-frequency spectrum; see Figs. 3 and 4. Our model predicts the data (Fig. 3). Under Kingman's coalescent (Fig. 4) one data point is outside the (2.5%, 97.5%) quantile-interval: the number of singleton polymorphisms. Interestingly, our model specifically predicts the one middle-frequency polymorphism (Fig. 3B), which exactly is the restriction-site polymorphism that distinguishes the two high-frequency haplotypes in the data.

Example 2. We used simulation Algorithm 1, described above, to show how the expected site-frequency spectrum under our model can differ from that under Kingman's coalescent. Recall that Algorithm 1 is for the case $c = \infty$ ($N^2\epsilon_N \to \infty$), in which case SREs dominate the ancestral process. We considered a number of different parameters values: $Y = 1, 2, 5, 10, 40, \infty$, and $\phi = 0.5$ ($Y = \infty$ corresponds to the Kingman's coalescent) and a sample size of 50. Fig. 5 shows the average site-frequency spectra for each set of parameters. The site-frequency spectra for $Y = 1$ and $Y = 2$ ($n = 50$) differ dramatically from what we expect under Kingman's coalescent. In particular, they are multi-modal, and are not decreasing, convex functions of the frequency as is the case in Kingman's coalescent. This surprising behavior is a consequence of there being multiple mergers in the ancestry; in Appendix B we prove that the site-frequency spectrum is a decreasing convex function of the frequency for the "general coalescent trees" of Griffiths and Tavaré (1998), which have only binary mergers but may have any distributions of coalescence times.

Next, we show that the presence of multiple mergers also affects the probability that an allele is the ancestral allele at a biallelic segregating site, as a function of its frequency. In the case where the expected site-frequency spectrum is a decreasing sequence, the major allele is more likely to be the ancestral allele, as it is the case for Kingman's coalescent and for general coalescent trees. For our model, this is not necessarily the case. In fact, we show below that cases exist in which an allele in low frequency is more likely to be the ancestral allele. It is also possible that alleles with different frequencies are equally likely to be the ancestral allele.

Using the results of Griffiths and Tavaré (1998) and Nielsen (2000), it follows that expected number $S(i)$ of mutant sites in frequency i in a sample of size n, is given by

$$S(i) = \frac{\theta}{2}\mathbb{E}L_i,$$

and the probability $P(i)$ that a mutant allele at the segregating site in a sample has frequency i is given by

$$\mathbb{P}(i) = \frac{\mathbb{E}\left(\frac{\mu}{2}L_i e^{-\frac{\mu}{2}L_n}\right)}{\mathbb{E}\left(\frac{\mu}{2}L_n e^{-\frac{\mu}{2}L_n}\right)},$$

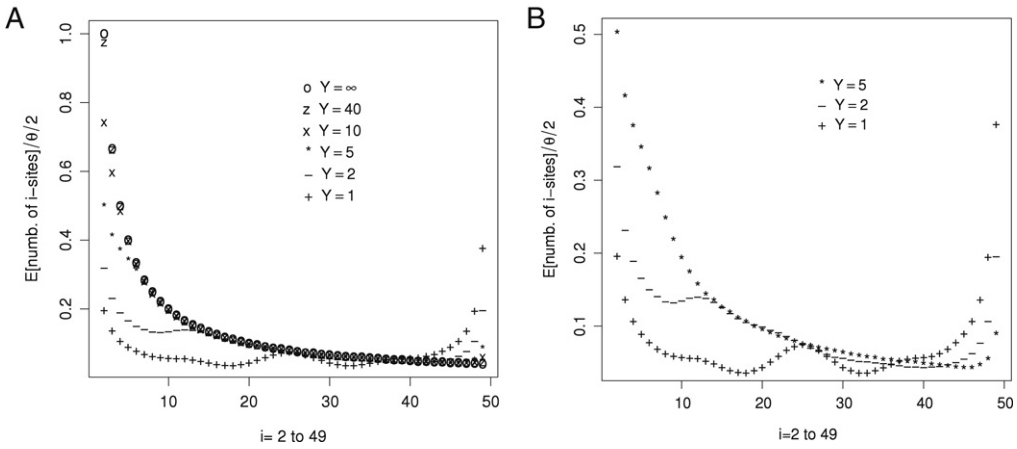

Fig. 5. Plots of expected site-frequency spectrum for sample size $n = 50$, $\phi = 0.5$, $Y = 1, 2, 5, 10, 40, \infty$, in (A). In (B) Y is equal to 1, 2, 5. In both plots, $c = \infty$. Here $Y = \infty$ gives Kingman's coalescent. The horizontal axis is the frequencies of the segregating sites that are greater than 1.

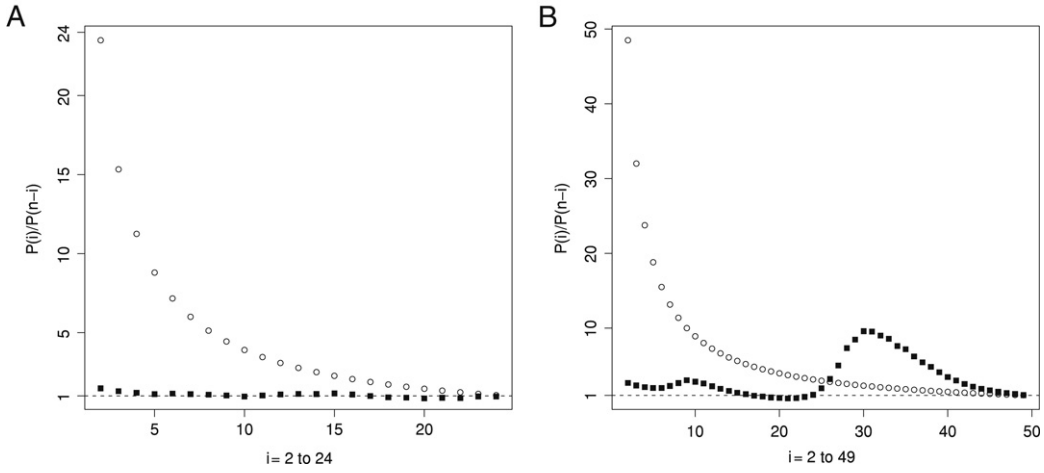

Fig. 6. The ratio, $P(i)/P(n-i)$, of probabilities of allele with frequency $n-i$ being ancestral allele versus allele with frequency $i = 2, \ldots, \lfloor (n-1)/2 \rfloor$. The points are marked by "■" and "○" respectively, for simultaneous multiple mergers and Kingman's coalescent cases. In (A) the sample size is 50, and the values of parameters are $Y = 1$ and $\phi = 0.5$. In (B) the sample size is 100, and the values of parameters are $Y = 2$ and $\phi = 0.7$.

where again L_i, $i = 1, \ldots, n-1$ is the sum of the lengths of the branches in the tree with i descendants in the sample, and where L_n is the total length of the tree. Then if we think of a nucleotide site as a very small locus, (specifically, $\mu \to 0$), we have

$$P(i) = \frac{\mathbb{E}L_i}{\mathbb{E}L_n}.$$

Therefore,

$$S(i) \propto P(i),$$

which we use below to estimate the ratios $P(i)/P(n-i)$.

Let A and a be two alleles at a biallelic segregating site with frequencies $f_A = i$ and $f_a = n-i$. We can use a Bayesian argument to predict which allele is more likely to be the ancestral allele. We denote the prior probabilities for allele A or a to be ancestral as P_A and P_a, respectively. The posterior probabilities can be expressed as

$$\mathbb{P}(A \mid f_A = i, f_a = n-i) = \frac{\mathbb{P}(f_A = i, f_a = n-i \mid A)P_A}{\mathbb{P}(f_A = i, f_a = n-i)}$$

and

$$\mathbb{P}(a \mid f_A = i, f_a = n-i) = \frac{\mathbb{P}(f_A = i, f_a = n-i \mid a)P_a}{\mathbb{P}(f_A = i, f_a = n-i)},$$

where

$$\mathbb{P}(f_A = i, f_a = n-i) = \mathbb{P}(f_A = i, f_a = n-i \mid a)P_a + \mathbb{P}(f_A = i, f_a = n-i \mid A)P_A.$$

By definition, we have

$$\mathbb{P}(f_A = i, f_a = n-i \mid a) = P(i)$$

and

$$\mathbb{P}(f_A = i, f_a = n-i \mid A) = P(n-i).$$

Assuming a uniform prior for each allele to be the ancestral allele, that is $P_a = P_A = 1/2$, then the ratio of the posterior probabilities is

$$\frac{\mathbb{P}(a \mid f_A = i, f_a = n-i)}{\mathbb{P}(A \mid f_A = i, f_a = n-i)} = \frac{P(i)}{P(n-i)} = \frac{S(i)}{S(n-i)}.$$

We have already observed using simulations (see Fig. 5) that $S(i)$ is not necessarily a decreasing function of i. We can use simulation results together with the above equation to estimate the ratios $P(i)/P(n-i)$, $i = 2, \ldots, n/2$.

To do this, we consider the case $c = \infty (N^2\epsilon_N \to \infty)$ and either a sample size of $n = 50$ with $Y = 1$ and $\phi = 0.5$; or a sample size of $n = 100$ with $Y = 2$ and $\phi = 0.7$. The estimated ratios $P(i)/P(n-i)$ are plotted in Fig. 6. For the comparison, we plotted the ratios for the case of Kingman's coalescent. Since for Kingman's coalescent case Fu (1995) has shown that $S(i) = \theta/i$ hence the ratios, $P(i)/P(n-i) = (n-i)/i, i = 1, \ldots, \lfloor (n-1)/2 \rfloor$, are always greater than one for Kingman's coalescent, which means that the major allele is more likely to be the ancestral allele. In the case of

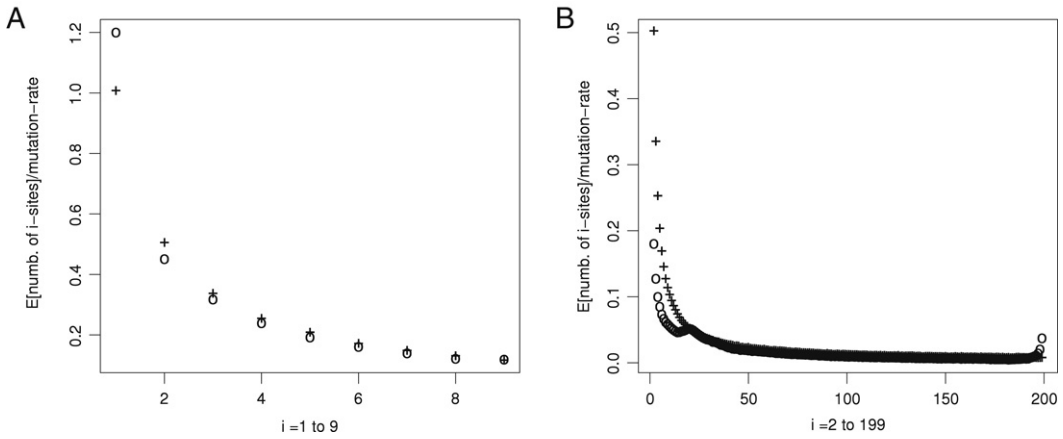

Fig. 7. In (A) and (B) the sample sizes are $n = 10$ and $n = 200$, respectively. The plots are the expected site-frequency spectrum under Kingman's coalescent and simultaneous multiple mergers divided by mutation rate, the points are marked respectively for each case by "+" and "o". The values of the parameters for simultaneous multiple mergers case are $\theta = 20, \phi = 0.5, Y = 5$; for Kingman's coalescent, the mutation rate is $\theta_K = 23$.

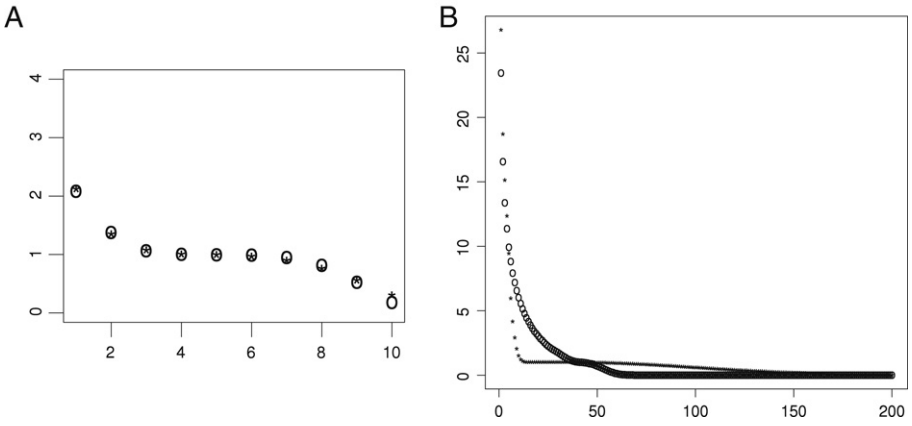

Fig. 8. The expected haplotype-frequency spectrum under Kingman's coalescent and simultaneous multiple mergers, with points for each case marked "o" and "*" respectively. In (A) and (B) the sample sizes are $n = 10$ and $n = 200$. In simultaneous multiple mergers case, the values of the parameters are: $\theta = 20, \phi = 0.5$, $Y = 5$. For Kingman's coalescent case the mutation rate, $\theta_K = 23$.

our model, one can see from Fig. 6, that these ratios, $P(i)/P(n-i)$, can be less than one or very close to one, meaning that the major allele is not always more likely to be the ancestral allele.

Example 3. In our final example, we show that for fixed parameter values in our model the difference between the predicted patterns of genetic variation under our model and under Kingman's coalescent can depend crucially on the sample size. To illustrate, we consider two values of the sample size: $n = 10$ and $n = 200$. The values for the parameters in our model are $Y = 10, \phi = 0.5$, $c = \infty$ and mutation rate $\theta = 20$. In this case, the ancestral process is dominated by SREs, and when an SRE occurs $\phi = 0.5$ of the population is replaced by the offspring of $Y = 10$ individuals. For Kingman's coalescent we used a mutation rate of $\theta_K = 23$, which was chosen to give roughly the same overall level of polymorphism in the two models.

We used simulations to estimate the expected site-frequency and haplotype-frequency spectrum under our model and under Kingman's coalescent. As one can see from Figs. 7 and 8, when the sample size is 10 the expected patterns of genetic variation in the sample (specifically, the expected haplotype-frequency and site-frequency spectra) from our model are very similar in shape to the expected patterns under Kingman's coalescent. However, when the sample size is 200, then the predictions of our model are significantly different than those of Kingman's coalescent. This implies that the power to reject the Kingman's coalescent may depend strongly on the sample size. We note that this effect depends on Y: the larger Y is the larger the sample size needs to be in order to detect a difference from Kingman's coalescent (results not shown).

3. Discussion

We have presented a forward-time reproduction model that captures life history characteristics common to many marine organisms. To understand patterns of genetic variation in a sample of DNA sequences, we studied a number of different coalescent approximations. Depending on the behavior of the underlying parameters in the model, we find three different types of ancestral processes: Kingman's coalescent, a coalescent with asynchronous multiple mergers, and a coalescent with simultaneous multiple mergers. The model we analyze here is a generalization of the model presented in Eldon and Wakeley (2006) in which $Y = 1$ and only asynchronous multiple mergers could occur. The model in Eldon and Wakeley (2006) is a generalization of Moran model, whereas our model represents a family of models that form a continuum between Wright-Fisher models and Moran models.

We have also included the possibility in our model that X_N and Y_N are random variables. For example, in the case of mussels living in the intertidal zone of the Pacific Northwest, Paine and Levin (1981) showed that the habitat patch sizes resulting from disturbance in winter storms, and other agents such as drifting

logs, show an approximations log-normal distribution of areas. Although we do not model spatial structure, this distribution of patch sizes corresponds to the distribution of X_N in our model.

To make some connection with other recent work on multiple-mergers coalescent processes, when $Y_N = 1$ and X_N has a distribution then our model can converge to the so-called Λ-coalescent processes, which has also been derived from branching processes (Birkner et al., 2005). Thus, our model is also similar to the family of coalescent processes currently being used to approximate the genealogy of a sample from a neutral locus that is linked to a locus that is under recurrent selective sweeps (Gillespie, 2000; Durrett and Schweinsberg, 2004; Schweinsberg and Durrett, 2005; Durrett and Schweinsberg, 2005).

As we have shown, mainly using simulations, commonly-held intuitions about genetic variation in a sample, which derive from Kingman's coalescent, may not apply in the case of multiple-mergers coalescent processes. First, our model does not always predict that an allele in high frequency in a sample is likely to be the oldest allele in the sample. This can be understood most simply in the case $Y = 2$, $\phi = 1$ and $c = \infty$, in which case the ancestry is dominated by rather extreme SREs. When the first SRE occurs in the ancestry of a sample of size n (and let us assume that no mutations occur during the waiting time to this event), then the numbers of descendants of the $Y = 2$ individuals are determined by a Binomial$(n, 1/2)$ distribution. It will be likely that one or the other of the two individuals will be the ancestor of the majority of the sample. The remainder of the ancestry is simply the waiting time back to a common ancestor of the two ancestral lineages. If a mutation occurs before this final coalescent event, it will be equally likely to occur on the branch ancestral to either of the $Y = 2$ individuals involved in the first (SRE) event. Thus, each allele will have a 50% chance of being the oldest allele, regardless of the frequencies.

We have also shown that when the sample size is small, it will likely be difficult to detect the signature of sweepstakes effect (Hedgecock, 1994), that is the presence of SREs. Some intuition can be gained from our analysis of the case $Y_N \to \infty$, where the ancestral process converges to Kingman's coalescent. Simply put, when the number of possible ancestors of the sample is very large, multiple-mergers coalescent events will be unlikely compared to binary mergers. Our simulations demonstrate the same kind of effect in the case where, technically, the ancestral process does not converge to Kingman's coalescent but rather to a coalescent with simultaneous multiple mergers ($Y > 1$, but finite). If the sample size is much smaller than Y, the probability of a binary merger will be roughly Y times more likely than the probability of a multiple merger. If only binary mergers happen to occur, then the ancestry will look very much like Kingman's coalescent. So, very large sample sizes may be required in order to detect the presence of SREs when there are a large number of potential parents at each SRE.

Appendix A

The proof of the fact that the approximation of the genealogy of a sample under our model is Kingman's coalescent if the following condition holds: $Y_N \to \infty$ or $X_N/N \to 0$.

According to Table 1, the genealogy of a sample can be approximated by Kingman's coalescent if the following limit conditions are satisfied:

$$\mathbb{P}_1(1,1) \to 0 \quad \text{and} \quad \frac{\mathbb{P}_1(1,1,1)}{\mathbb{P}_1(1,1)} \to 0.$$

First, from approximation (4) it follows that $\mathbb{P}_1(1,1) \to 0$ as $Y_N \to \infty$ or $X_N/N \to 0$. Second, because of the approximations (4)–(5) the second limit condition is equivalent to the condition that the ratio $\frac{\tilde{\mathbb{P}}_1(1,1,1)}{\tilde{\mathbb{P}}_1(1,1)}$ converges to zero as $N \to \infty$, where $\tilde{\mathbb{P}}_1(1,1,1)$ and $\tilde{\mathbb{P}}_1(1,1)$ are the expressions on the right in the approximations (4), (5), respectively. And to prove it, first, we get an upper bound for the ratio $\frac{\tilde{\mathbb{P}}_1(1,1,1)}{\tilde{\mathbb{P}}_1(1,1)}$ then we show that the upper bound becomes small as $N \to \infty$.

One can easily see that

$$\epsilon_N \mathbb{E}\frac{X_N^2}{N^2}\left(\frac{X_N-1}{X_N Y_N} + \frac{2(1-\frac{X_N}{N})}{X_N}\right) \le \tilde{\mathbb{P}}_1(1,1).$$

Hence we have

$$\frac{\tilde{\mathbb{P}}_1(1,1,1)}{\tilde{\mathbb{P}}_1(1,1)} \le \frac{\epsilon_N \mathbb{E}\frac{X_N^2(X_N-1)}{N^3 Y_N}\left(\frac{X_N-2}{X_N Y_N} + \frac{3(1-\frac{X_N}{N})}{X_N}\right)}{\epsilon_N \mathbb{E}\frac{X_N^2}{N^2}\left(\frac{X_N-1}{X_N Y_N} + \frac{2(1-\frac{X_N}{N})}{X_N}\right)}.$$

It is obvious that

$$\frac{X_N-2}{X_N Y_N} + \frac{3(1-\frac{X_N}{N})}{X_N} \le \frac{3}{2}\left(\frac{X_N-1}{X_N Y_N} + \frac{2(1-\frac{X_N}{N})}{X_N}\right),$$

therefore we have

$$\frac{\mathbb{E}\frac{X_N^2}{N^2}\left(\frac{X_N-2}{X_N Y_N} + \frac{3(1-\frac{X_N}{N})}{X_N}\right)}{\mathbb{E}\frac{X_N^2}{N^2}\left(\frac{X_N-1}{X_N Y_N} + \frac{2(1-\frac{X_N}{N})}{X_N}\right)} \le 3/2.$$

Now, because of the condition: $Y_N \to \infty$ or $X_N/N \to 0$, it follows that $\frac{(X_N-1)}{NY_N}$ uniformly converges to zero, which means that for any positive δ the ratio $\frac{(X_N-1)}{NY_N}$ is less than δ for big N (more precisely: there is an N_δ such that the inequality holds if $N > N_\delta$). Hence,

$$\frac{\epsilon_N \mathbb{E}\left(\frac{X_N^2(X_N-1)}{N^3 Y_N}\left(\frac{X_N-2}{X_N Y_N} + \frac{3(1-\frac{X_N}{N})}{X_N}\right)\right)}{\epsilon_N \mathbb{E}\left(\frac{X_N^2}{N^2}\left(\frac{X_N-1}{X_N Y_N} + \frac{2(1-\frac{X_N}{N})}{X_N}\right)\right)} < \delta \frac{\mathbb{E}\frac{X_N^2}{N^2}\left(\frac{X_N-2}{X_N Y_N} + \frac{3(1-\frac{X_N}{N})}{X_N}\right)}{\mathbb{E}\frac{X_N^2}{N^2}\left(\frac{X_N-1}{X_N Y_N} + \frac{2(1-\frac{X_N}{N})}{X_N}\right)},$$

as $N > N_\delta$. Thus, combining the above estimates we have

$$\frac{\tilde{\mathbb{P}}_1(1,1,1)}{\tilde{\mathbb{P}}_1(1,1)} < \frac{3}{2}\delta, \quad \text{as } N > N_\delta.$$

Therefore, the ratio $\frac{\tilde{\mathbb{P}}_1(1,1,1)}{\tilde{\mathbb{P}}_1(1,1)}$ converges to zero because the right-hand side of the inequality above can be made arbitrarily small.

Appendix B

The proof of the fact that the probability distribution of the frequency of a mutant allele in the sample is a decreasing and convex sequence when the genealogies of samples are approximated by general coalescent trees.

For the cases when the genealogies of samples are approximated by general coalescent trees (see Griffiths and Tavaré (1998, 2003)) the probability distribution of the frequency of a mutant allele at a particular segregating site in a sample is

$$P(i) = \frac{\sum_{k=2}^{n} k p_{n,k}(i)\mathbb{E}T_k}{\sum_{k=2}^{n} k\mathbb{E}T_k}$$

where $P(i)$ is the probability that the mutant allele at a segregating site (in a sample of size n) has frequency i; T_k is the waiting time to the next coalescent event (looking backwards in time) in a

general coalescent tree when the sample has k ancestors just after a coalescent event; $p_{n,k}(i)$, $i = 1, \ldots, n-1$ are defined as follows:

$$p_{n,k}(i) = \begin{cases} \dfrac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} & \text{if } 2 \le k \le n-i+1, \\ 0 & \text{if } k > n-i+1. \end{cases} \tag{12}$$

We assume that $\mathbb{E}T_k \neq 0$, $k = 2, \ldots, n$ and $n > 2$. Under these conditions, first we show that this probability distribution is a deceasing and convex sequence. However, if these conditions do not hold then from the proof below, it follows that the probability density is a non-increasing sequence.

First we show that $p_{n,k}(b)$ is decreasing with respect to b, $b = 2, \ldots, n-k+1$, for fixed n and k, $k > 2$. From the definition of $p_{n,k}(i)$ one can easily check that the following identity is true:

$$p_{n,k}(i) - p_{n,k}(i+1) = p_{n,k}(i)\frac{k-2}{n-i-1},$$

if $1 \le i \le n-k$. From this identity it follows that, for fixed n and k, $p_{n,k}(i)$ is a decreasing sequence with respect to i, $i = 1, \ldots, n-k$, when $k > 2$; if $k = 2$ then $p_{n,k}(i)$ is a constant with respect to i. Also, it is obvious that $p_{n,k}(i) - p_{n,k}(i+1) \ge 0$ for i, $n-k < i \le n-1$. Thus, $p_{n,k}(i)$ is a non-increasing sequence for $i = 1, \ldots, n-1$, but it is a decreasing sequence if $k = 3$.

We have that $P(i)$ is the linear combination of non-increasing sequences, $p_{n,k}(i)$, with non-negative coefficients, therefore $P(i)$ is non-increasing. Since we assume that $\mathbb{E}T_3 \neq 0$ and $p_{n,3}(i)$ is a decreasing sequence, therefore $P(i)$ is a decreasing sequence.

Now, to show that $P(i)$ is a convex sequence we check that the sequences $p_{n,k}(i)$ are convex with respect to i, $i = 1, \ldots, n-1$. Thus, it is enough to see that the following conditions hold (for fixed n and k): $2p_{n,k}(i) - (p_{n,k}(i-1) + p_{n,k}(i+1)) \le 0$, where $i = 2, \ldots, n-2$. From the definition of $p_{n,k}(i)$ one can easily see that the following identities are true:

$$2p_{n,k}(i) - (p_{n,k}(i-1) + p_{n,k}(i+1)) = -p_{n,k}(i)\frac{(k-2)(k-3)}{(n-i-1)(n-i-k+2)},$$

if $n-k+1 < i \le n-1$; and

$$2p_{n,k}(n-k+1) - p_{n,k}(n-k) = \frac{(3-k)}{\binom{n-1}{k-1}},$$

if $k > 2$. Since from definition (12) we have $p_{n,k}(j) = 0$, $n-k+1 < j \le n-1$, combining this with the above identities we have that $p_{n,k}(i)$ is a convex sequence. Because $P(i)$ is the linear combination of $p_{n,k}(i)$ with non-negative coefficients, therefore $P(i)$ is a convex sequence.

References

Beckenbach, A.T., 1994. Mitochondrial haplotype frequencies in oyster: Neutral alternatives to selection models. In: Gelding, B. (Ed.), Non Neutral Evolution: Theories and Molecular Data. Chapman and Hall, New York, NJ, pp. 188–198.

Birkner, M., Blath, J., Capaldo, M., Etheridge, A., Möhle, M., Schweinsberg, J., Wakolbinger, A., 2005. Alpha-stable branching processes and beta-coalescents. Electronic Journal of Probability 10, 303–325.

Boom, J.D.G., Boulding, E.G., Beckenbach, A.T., 1994. Mitochondrial DNA variation in introduced populations of Pacific oyster, Crassostrea gigas, in British Columbia. Canadian Journal of Fisheries and Aquatic Sciences 51, 1608–1614.

Cannings, C., 1973. The equivalence of some overlapping and non-overlapping generation models for the study of genetic drift. Journal of Applied Probability 10, 432–436.

Cannings, C., 1974. The latent roots of certain Markov chains arising in genetics: A new approach, I. Haploid models. Advances in Applied Probability 6, 260–290.

Cannings, C., 1975. The latent roots of certain Markov chains arising in genetics: A new approach, II. Further happloid models. Advances in Applied Probability 7, 264–282.

Dayton, P.K., 1971. Competition, disturbance, and community organization: The provision and subsequent utilization of space in a rocky intertidal community. Ecological Memographs 41, 351–389.

Durrett, R., Schweinsberg, J., 2004. Approximating selective sweeps. Theoretical Population Biology 66, 129–138.

Durrett, R., Schweinsberg, J., 2005. A coalescent model for the effect of advantageous mutations on the genealogy of a population. Stochastic Processes and their Applications 115, 1628–1657.

Eldon, B., Wakeley, J., 2006. Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics 172, 2621–2633.

Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theoretical Population Biology 3, 87–112.

Flowers, J.M., Schroeter, S.C., Burton, R.S., 2002. The recruitment sweepstakes has many winners: Genetic evidence from the sea urchin Strongylocentrotus purpuratus. Evolution 56, 1445–1453.

Fu, Y.-X., 1995. Statistical properties of segregating sites. Theoretical Population Biology 48, 172–197.

Gillespie, J.H., 2000. Genetic drift in an infinite population. The pseudohitchhiking model. Genetics 155, 909–919.

Griffiths, R.C., Tavaré, S., 1998. The age of a mutation in a general coalescent tree. Communications in Statistics-Stochastic Models 14, 273–295.

Griffiths, R.C., Tavaré, S., 2003. The genealogy of a neutral mutation. In: Green, P.J., Hjort, N.L., Richardson, S. (Eds.), Highly Structured Stochastic Systems. Oxford Statistical Science. Oxford University Press, Oxford, pp. 393–413.

Hedgecock, D., 1994. Does variance in reproductive success limit effective population sizes of marine organisms? In: Beaumont, A.R. (Ed.), Genetics and Evolution of Aquatic Organisms. Chapman and Hall, London, UK, pp. 122–134.

Hedrick, P., 2005. Large variance in reproductive success and the N_e/N ratio. Evolution 59, 1596–1599.

Kingman, J.F.C., 1982a. On the genealogy of large populations. Journal of Applied Probability 19A, 27–43.

Kingman, J.F.C., 1982b. The coalescent. Stochastic Processes and their Applications 13, 235–248.

Möhle, M., Sagitov, S., 2001. A classification of coalescent processes for haploid exchangeable population models. The Annals of Probability 29, 1547–1562.

Nielsen, R., 2000. Estimating of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154, 931–942.

Paine, R.T., Levin, S.A., 1981. Intertidal landscapes: Disturbance and the dynamics of pattern. Ecological Memographs 51, 145–178.

Pitman, J., 1999. Coalescents with multiple collisions. Annals of Probability 27, 1870–1902.

Reeb, C.A., Avise, J.C., 1990. A genetic discontinuity in a continuously distributed species: Mitochondrial DNA in the American oyster, *Crassostrea virginica*. Genetics 124, 397–406.

Sagitov, S., 1999. The general coalescent with asynchronous mergers of ancestral lines. Journal of Applied Probability 36, 1116–1125.

Sagitov, S., 2003. Convergence to the coalescent with simultaneous mergers. Journal of Applied Probability 40, 839–854.

Schweinsberg, J., 2000. Coalescents with simultaneous multiple collisions. Electronic Journal of Probability 5, 1–50.

Schweinsberg, J., Durrett, R., 2005. Random partitions approximating the coalescence of lineages during a selective sweep. Annals of Applied Probability 15, 1591–1651.

Wakeley, J., Takahashi, T., 2003. Gene genealogies when the sample size exceeds the effective size of the population. Molecular Biology and Evolution 20, 208–213.

Witman, J.D., 1987. Subtidal coexistence: Storms, grazing, mutualism, and the zonation of kelps and mussels. Ecological Memographs 57, 167–187.