Robust model selection between population growth and multiple merger coalescents

Jere Koskela
j.koskela@warwick.ac.uk
Department of Statistics
University of Warwick
Coventry, CV4 7AL
United Kingdom

Maite Wilke Berenguer
maite.wilkeberenguer@ruhr-uni-bochum.de
Fakultät für Mathematik
Ruhr Universität Bochum
Universitätstraße 150, 44780 Bochum,
Germany

March 5, 2019

Abstract

We study the effect of biological confounders on the model selection problem between Kingman coalescents with population growth, and Ξ-coalescents involving simultaneous multiple mergers. We use a low dimensional, computationally tractable summary statistic, dubbed the *singleton-tail statistic*, to carry out approximate likelihood ratio tests between these model classes. The singleton-tail statistic has been shown to distinguish between them with high power in the simple setting of neutrally evolving, panmictic populations without recombination. We extend this work by showing that cryptic recombination and selection do not diminish the power of the test, but that misspecifying population structure does. Furthermore, we demonstrate that the singleton-tail statistic can also solve the more challenging model selection problem between multiple mergers due to selective sweeps, and multiple mergers due to high fecundity with moderate power of up to 30%.

1 Introduction

The Kingman coalescent [Kingman, 1982a,b,c, Hudson, 1983a,b, Tajima, 1983] models ancestral relations of samples from large populations as random, binary trees, and is an important tool for predicting genetic diversity. A central assumption of the Kingman coalescent is low variance of family sizes, so that large populations always consist of many relatively small families. Violations of this assumption call for models with infinite variance family sizes, and lead to so called Λ-coalescents, which allow more than two lineages to merge to a common ancestor simultaneously [Donnelly and Kurtz, 1999a, Pitman, 1999, Sagitov, 1999].

There is growing evidence that Λ-coalescents are an appropriate model for organisms with high fecundity coupled with a skewed offspring distribution [Beckenbach, 1994, Árnason, 2004, Eldon and Wakeley, 2006, Sargsyan and Wakeley, 2008, Hedgecock and Pudovkin, 2011, Birkner et al., 2011, Steinrücken et al., 2013, Tellier and Lemaire, 2014]. Consequently, development of statistical techniques for distinguishing the Kingman coalescent from Λ-coalescents has also been an active area of research; see [Eldon et al., 2015, Koskela, 2018], and references therein. In particular, attention has focused on distinguishing Λ-coalescents from Kingman coalescents with population growth, because both classes of models predict an excess of singletons (mutations only carried by one individual in a sample of DNA sequences) relative to the standard Kingman coalescent under the infinitely many sites model of mutation [Watterson, 1975].

Koskela [2018] introduced a simple, two-dimensional summary statistic, referred to here as the *singleton-tail statistic*, which distinguishes between these model classes with high power even from a data set consisting of 500 samples from bi-parental, diploid organisms sequenced at around 10 unlinked chromosomes. The correct model could be selected with high power without knowing the population-rescaled mutation rate, provided it is was not very low (see also [Eldon et al., 2015, Supporting Information 12]). In this paper we investigate the impact of other confounders on the prospect of discriminating between these models based on the singleton-tail statistic, again in the bi-parental, diploid setting. In particular, we will focus on each of

1. weak natural selection modelled by the Ancestral Selection Graph [Krone and Neuhauser, 1997, Neuhauser and Krone, 1997, Donnelly and Kurtz, 1999b, Baake et al., 2016],
2. crossover recombination within chromosomes modelled by the Ancestral Recombination Graph [Hudson, 1983a, Griffiths and Marjoram, 1997, Donnelly and Kurtz, 1999b, Birkner et al., 2013],
3. population structure modelled by the structured coalescent [Herbots, 1997, Limic and Sturm, 2006, Eldon, 2009].

We will demonstrate that the presence or absence of the first two has minimal effect on the performance of the hypothesis test developed in [Koskela, 2018], while population structure is a significant counfounder that must be correctly incorporated into the model.

There are four parental copies of each chromosome involved in each merger in the diploid, bi-parental setting, allowing for up to four simultaneous mergers. Hence the models considered in this paper are actually Ξ-coalescents [Schweinsberg, 2000, Möhle and Sagitov, 2001] which allow simultaneous multiple mergers, despite the fact that the population will be assumed to reproduce in a fashion consistent with the more restrictive Λ-coalescent permitting only one multiple merger at a time.

We also use the singleton-tail statistic to distinguish two classes of Λ-coalescents: those arising from high fecundity reproduction, and those arising from selective sweeps [Durrett and Schweinsberg, 2005]. This problem is more challenging than a null hypothesis consisting of Kingman coalescents with population growth, because the marginal coalescent process at each chromosome can be identical under the two hypotheses. However, high fecundity reproduction results in positively correlated coalescence times between unlinked chromosomes, whereas unlinked chromosomes are independent under the selective sweep model. The positive correlation results in increased sampling variance of the singleton-tail statistic, which yields tests with moderate statistical power of up to 30%.

The rest of the paper is organised as follows. In Section 2 we recall the singleton-tail statistic of [Koskela, 2018] as well as the associated hypothesis test for model selection. Section 3 presents a unified, diploid coalescent model incorporating high fecundity reproduction and population growth, as well as the three confounders of weak selection, crossover recombination, and discrete spatial structure. Models with only population growth or high fecundity reproduction, as well as any desired subset of confounders, can be recovered as special cases. Section 4 provides simulation studies on the effect of each of the three confounders on the sampling distribution of the singleton-tail statistic, as well as the associated hypothesis test. In Section 5 we introduce a different model in which rapid selective sweeps result in multiple mergers acting locally on the genome, and investigate whether the singleton-tail statistic can distinguish it from the Ξ-coalescent introduced in Section 3. Section 6 concludes with a discussion.

2 The singleton-tail statistic

Suppose a sample of $n \in \mathbb{N}$ DNA sequences from a single chromosome is available, and that derived mutations can be distinguished from ancestral states. Let $[n] := \{1, \ldots, n\}$, and let $\xi_i^{(n)}$

be the number of sites at which a mutant allele appears $i \in [n-1]$ times. Then

$$\boldsymbol{\xi}^{(n)} := \left(\xi_1^{(n)}, \ldots, \xi_{n-1}^{(n)}\right)$$

is the *unfolded site-frequency spectrum* (SFS). If mutant and ancestral types cannot be distinguished, the *folded* spectrum $\boldsymbol{\eta}^{(n)} := (\eta_1^{(n)}, \ldots, \eta_{\lfloor n/2 \rfloor}^{(n)})$ [Fu, 1995] is used instead, where

$$\eta_i^{(n)} := \frac{\xi_i^{(n)} + \xi_{n-i}^{(n)}}{1 + \delta_{i,n-i}}, \quad 1 \leq i \leq \lfloor n/2 \rfloor,$$

and $\delta_{i,j} = 1$ if $i = j$, and is zero otherwise. Let $\boldsymbol{\zeta}^{(n)} := (\zeta_1^{(n)}, \ldots, \zeta_{n-1}^{(n)})$ be the normalised unfolded SFS, whose entries are given by $\zeta_i^{(n)} := \xi_i^{(n)} / |\boldsymbol{\xi}^{(n)}|$, where $|\boldsymbol{\xi}^{(n)}| := \xi_1^{(n)} + \cdots + \xi_{n-1}^{(n)}$ is the total number of segregating sites, and with the convention that $\boldsymbol{\zeta}^{(n)} = \mathbf{0}$ if there are no segregating sites.

Now, for any $k \in [n-1]$ define the *lumped tail* of the SFS as

$$\overline{\zeta}_k^{(n)} := \sum_{j=k}^{n-1} \zeta_j^{(n)},$$

and consider the summary statistic $(\zeta_1^{(n)}, \overline{\zeta}_k^{(n)})$ for some fixed k. Data from multiple chromosomes is incorporated by averaging: if L unlinked chromosomes are available, then the singleton-tail statistic is

$$(\zeta_{1,L}^{(n)}, \overline{\zeta}_{k,L}^{(n)}) := \frac{1}{L} \sum_{j=1}^{L} (\zeta_1^{(n)}(j), \overline{\zeta}_k^{(n)}(j)),$$

where $(\zeta_1^{(n)}(j), \overline{\zeta}_k^{(n)}(j))$ denotes the singleton class and lumped tail computed from the j^{th} chromosome.

For two classes of models Θ_0 and Θ_1, the likelihood ratio test statistic is

$$\frac{\sup_{\Pi \in \Theta_1} P^{\Pi}(\zeta_{1,L}^{(n)}, \overline{\zeta}_{k,L}^{(n)})}{\sup_{\Pi \in \Theta_0} P^{\Pi}(\zeta_{1,L}^{(n)}, \overline{\zeta}_{k,L}^{(n)})},$$

where P^{Π} denotes the sampling distribution of the singleton-tail statistic under coalescent Π. A corresponding hypothesis test of size $\omega \in (0, 1)$ given an observed value of the singleton-tail statistic is

$$\Phi(\zeta_{1,L}^{(n)}, \overline{\zeta}_{k,L}^{(n)}) = \begin{cases} 0 & \text{if } \frac{\sup_{\Pi \in \Theta_1} P^{\Pi}(\zeta_{1,L}^{(n)}, \overline{\zeta}_{k,L}^{(n)})}{\sup_{\Pi \in \Theta_0} P^{\Pi}(\zeta_{1,L}^{(n)}, \overline{\zeta}_{k,L}^{(n)})} \leq q_\omega \\ 1 & \text{if } \frac{\sup_{\Pi \in \Theta_1} P^{\Pi}(\zeta_{1,L}^{(n)}, \overline{\zeta}_{k,L}^{(n)})}{\sup_{\Pi \in \Theta_0} P^{\Pi}(\zeta_{1,L}^{(n)}, \overline{\zeta}_{k,L}^{(n)})} > q_\omega \end{cases}, \tag{1}$$

where $\Phi(\zeta_{1,L}^{(n)}, \overline{\zeta}_{k,L}^{(n)}) = 1$ corresponds to rejecting the null hypothesis Θ_0, and q_ω is the quantile

$$q_\omega := \inf \left\{ q \geq 0 : \sup_{\Pi \in \Theta_0} \mathbb{P}^{\Pi} \left(\frac{\sup_{\Pi \in \Theta_1} P^{\Pi}(\zeta_{1,L}^{(n)}, \overline{\zeta}_{k,L}^{(n)})}{\sup_{\Pi \in \Theta_0} P^{\Pi}(\zeta_{1,L}^{(n)}, \overline{\zeta}_{k,L}^{(n)})} \geq q \right) \leq \omega \right\}.$$

The sampling distribution P^{Π} and the quantile q_ω are both intractable, but can easily be approximated by simulation due to the low dimensionality of the singleton-tail statistic to obtain an implementable hypothesis test with approximate size ω Koskela [2018]. In particular, we consider the hypotheses

$$\begin{aligned}
\Theta_0 := \{&\text{Kingman coalescent with exponential growth at population-rescaled rate} \\
&\gamma \in \{0, 0.1, 0.2, \ldots, 0.9, 1, 1.25, 1.5, 2, 2.5, 3, 3.5, 4, 5, 6, \ldots, 19, 20, \\
&\quad 25, 30, 35, 40, 50, 60, \ldots, 990, 1000\}\}, \\
\Theta_1 := \{&\text{Beta}(2-\alpha, \alpha)\text{-}\Xi\text{-coalescents with } \alpha \in \{1, 1.025, \ldots, 1.975, 2\}\}, \qquad (2)
\end{aligned}$$

In brief, data is simulated under both Θ_0 and Θ_1, and kernel density estimates (KDEs) $\hat{P}^{\Pi}$ of the intractable sampling distributions P^{Π} are obtained for $\Pi \in \Theta_0$ and $\Pi \in \Theta_1$. These KDEs, along with more simulated data, can be used to accurately approximate the intractable quantile q_ω, yielding an implementable hypothesis test. Our KDEs were obtained using the `kde` function in the `ks` package (version 1.10.4) in `R` under default settings. In particular, this method uses truncated Gaussian kernels, and determines bandwidths using the SAMSE estimator [Duong and Hazelton, 2003, equation (6)].

Remark 1. The null hypothesis in Koskela [2018] was broader and included algebraic population growth, in addition to exponential. However, results in Koskela [2018] showed that the two growth models resulted in very similar sampling distributions for the singleton-tail statistic, and hence we focus on the exponential growth model.

Simulating data in order to approximate the test (1) requires specification of the cutoff k for the lumped tail of the singleton-tail statistic, as well as of the mutation rate θ. Sensitivity analyses conducted in Koskela [2018] showed that the test was highly insensitive to the choice of k provided $k \gtrapprox 6$, as well as to misspecification of the mutation rate by up to a factor of ten. We fix $k = 15$ throughout, and use the known, true mutation rate in our simulation studies. For biological data sets with an unknown mutation rate, the analysis in Koskela [2018] demonstrated that it is sufficient to use the generalised Watterson estimator,

$$\hat{\theta} = |\boldsymbol{\xi}^{(n)}| / \mathbb{E}^{\Pi}[T^{(n)}],$$

where $\mathbb{E}^{\Pi}[T^{(n)}]$ is the expected branch length from n leaves under coalescent Π.

3 An umbrella model

In this section we describe a general class of models incorporating diploidy, bi-parental, high fecundity reproduction, population growth, weak natural selection, population structure in a discrete geography, and crossover recombination. This generalises both the Ancestral Influence Graph Donnelly and Kurtz [1999b], as well as time-inhomogeneous multiple merger coalescents Möhle [2002], Matuszewski et al. [2018]. Models with any subset of the above forces can be recovered as special cases.

Consider a geography of D demes, with the population size on deme $i \in [D]$ at time t given by $2M_N^{(i)}(t)$, where N is a scaling parameter. We also define the total population size

$$2M_N(t) := \sum_{i=1}^{D} 2M_N^{(i)}(t),$$

and the shorthand $M_N := M_N(0)$.

Each individual carries a diploid genome consisting of $L \in \mathbb{N}$ pairs of unlinked chromosomes. Each chromosome carries one of $K \in \mathbb{N}$ alleles, identified with $[K]$, which are acted upon by natural selection. In addition, each chromosome is identified with the unit interval $[0, 1]$, on which neutral mutations and crossover recombination take place. For definiteness, we assume that the selective allele is fully linked to the left end of the neutral interval.

The populations evolve in discrete time with non-overlapping generations. At each time t, the individuals in each deme form pairs uniformly at random. The pairs are ordered in a fixed but arbitrary way, and pair j on deme i has a random number of offspring denoted by $\nu_j^{(i)}(t) + \beta_j^{(i)}(t)$. The two summands will be associated with neutral reproduction and natural selection, respectively. As such, the distribution of $\beta_j^{(i)}(t)$ will depend on the alleles of the two parents, though this dependence is suppressed for legibility. Likewise, we will frequently suppress the time-dependence in the family sizes, and write $\nu_j^{(i)}$ and $\beta_j^{(i)}$. For future convenience, we define $\tilde{\beta}_j^{(i)}(t)$ as the random number of selective offspring that pair j on deme i at time t would have had if they carried the fittest possible combination of alleles.

The neutral offspring vectors $(\nu_1^{(i)}, \dots, \nu_{M_N^{(i)}(t)}^{(i)})$ are assumed to be exchangeable, and independent across demes as well as time steps. The selective offspring vectors $(\beta_1^{(i)}, \dots, \beta_{M_N^{(i)}(t)}^{(i)})$ are independent across demes and time steps. In addition, both vectors on each deme are assumed to satisfy the almost sure constraint

$$\sum_{j=1}^{M_N^{(i)}(t)} \nu_j^{(i)} + \beta_j^{(i)} \equiv 2M_N^{(i)}(t+1).$$

Each offspring inherits one copy of each of its L chromosome pairs from each of its parents. Each inherited chromosome is a mosaic of the two chromosomes carried by the parent, with the number of recombination breakpoints having the Poisson distribution with parameter r_N, and each break point being uniformly distributed along the chromosome. All of the Poisson and uniform random variables are independent of each other, as well as of the wider reproduction mechanism. Each locus inherits its allele from the parental chromosome assigned to its leftmost segment, with selective mutations happening independently at random with probability μ_N. Mutant types are drawn from a stochastic matrix $M = (M_{ij})_{i,j=1}^K$, where M_{ij} is the probability of a mutant locus having allele j given its parent had allele i.

After the reproduction step is complete, a deterministic fraction $m_{ij}^{(N)}$ of children chosen uniformly at random from deme i migrate to deme j, for each pair of demes. These migration fractions are assumed to satisfy

$$M_N^{(i)}(t) \sum_{j=1}^{D} m_{ij}^{(N)} \equiv \sum_{j=1}^{D} M_N^{(j)}(t) m_{ji}^{(N)}, \tag{3}$$

for each $t \geq 0$, so that the population sizes of demes remain unchanged by migration. For notational brevity we set $m_{ii}^{(N)} \equiv 0$.

We now reverse the direction of time, so that time $t \in \mathbb{N}$ corresponds to t generations in the past in the model specified above. For $n \in \mathbb{N}$ and $k \in \mathbb{N}$ let $(n)_k := n(n-1)\dots(n-k+1)$ denote the falling factorial, and define

$$\begin{aligned} c_N^{(i)}(t) &:= \frac{M_N^{(i)}(t)}{4(2M_N^{(i)}(t-1))_2} \mathbb{E}[(\nu_1^{(i)})_2], \\ c_N &:= \frac{1}{4(2M_N)_2} \sum_{i=1}^{D} M_N^{(i)}(1) \mathbb{E}[(\nu_1^{(i)})_2], \end{aligned} \tag{4}$$

as the probability that two chromosomes sampled uniformly at random from deme $i \in [D]$ (resp. the whole population) at time $t \in \mathbb{N}$ (resp. $t = 0$) were born to a common family in the previous generation, made the same choice from two available parents, and also the same choice of chromosome within that parent. In other words, $c_N^{(i)}(t)$ is the probability of two time t chromosomes on island i merging to a common ancestor in one generation, while c_N is the same probability for two chromosomes sampled uniformly from the whole population at time $t = 0$.

We make the following assumptions for each $i, j \in [D]$, and each $t \in (0, \infty)$, as $N \to \infty$, where each Λ_i is a probability measure on $[0, 1]$, and each $\lambda_i(t)$ is a positive function bounded away from 0, with $\lambda_1(0) + \ldots + \lambda_D(0) = 1$ and $\inf_{t \geq 0, i \neq j}\{\lambda_i(t)/\lambda_j(t)\} > 0$, and $\gamma \in [0, 1)$ and $C > 0$ are constant independent of N, i, and t:

$$c_N \to 0, \tag{5}$$

$$\inf_{i \in [D], t \geq 0}\{M_N^{(i)}(t)\} \to \infty, \tag{6}$$

$$\mathbb{E}[(\nu_1^{(i)}(t))_2] \sim C M_N^{(i)}(t)^{\gamma}, \tag{7}$$

$$\frac{M_N^{(i)}(\lfloor t/c_N \rfloor)}{M_N} \to \lambda_i(t), \tag{8}$$

$$\frac{(M_N^{(i)}(\lfloor t/c_N \rfloor))_2 \mathbb{E}[(\nu_1^{(i)} + \tilde{\beta}_1^{(i)})_2 (\nu_2^{(i)} + \tilde{\beta}_2^{(i)})_2]}{(2M_N^{(i)}(\lfloor t/c_N \rfloor - 1))_4 c_N} \to 0, \tag{9}$$

$$\frac{M_N^{(i)}(\lfloor t/c_N \rfloor - 1)}{c_N^{(i)}(\lfloor t/c_N \rfloor)} \mathbb{P}(\nu_1^{(i)} > 2M_N^{(i)}(\lfloor t/c_N \rfloor - 1)x) \to \int_x^1 \frac{\Lambda_i(dy)}{y^2}, \tag{10}$$

$$\mu_N/c_N \to \theta \in [0, \infty), \tag{11}$$

$$r_N/c_N \to \rho \in [0, \infty), \tag{12}$$

$$m_{ij}^{(N)}/c_N \to m_{ij} \in [0, \infty), \tag{13}$$

$$\mathbb{E}[\tilde{\beta}_1^{(i)}]/c_N \to \sigma_i \in [0, \infty), \tag{14}$$

$$\frac{1}{c_N} \sup_{k \geq 1} \left\{ \mathbb{E}\left[\tilde{\beta}_1^{(i)} \left(\nu_1^{(i)} + \sum_{j=1}^{M_N^{(i)}(\lfloor t/c_N \rfloor)} \tilde{\beta}_j^{(i)} \right)^k \right] \right\} \to 0. \tag{15}$$

Remark 2. It is well known that if $\Lambda_i = \delta_0$, the Dirac delta-measure at 0, in (10), then the assumption is equivalent to

$$\frac{\mathbb{E}[(\nu_1^{(i)})_3]}{4(2M_N^{(i)}(\lfloor t/c_N \rfloor - 1))_2 c_N^{(i)}(\lfloor t/c_N \rfloor)} \sim \frac{\left(\sum_{j=1}^D \lambda_j(0)^{\gamma+1}\right) \mathbb{E}[(\nu_1^{(i)})_3]}{16 M_N^2 \lambda_i(t)^{\gamma+1} c_N} \to 0$$

for each $t \in (0, \infty)$ and $i \in [D]$, where the second representation follows from (8) and

$$c_N^{(i)}(\lfloor t/c_N \rfloor) \sim \frac{\lambda_i(t)^{\gamma-1}}{\sum_{j=1}^D \lambda_j(0)^{\gamma+1}} c_N, \tag{16}$$

itself a consequence of (7) and (8). See [Möhle and Sagitov, 2003, Section 5] for details. This assumption disallows multiple mergers in the limiting ancestry, which will thus only consist of isolated binary mergers. Any other choice of Λ_i will yield an ancestry with up to four simultaneous multiple mergers at each chromosome, corresponding to the four possible parental chromosomes involved in the forwards-in-time reproduction event, and thus produce ancestries described by a Ξ-coalescent. See [Möhle and Sagitov, 2003, Section 6] for details of Ξ-coalescents arising out of diploid reproduction in this way.

Remark 3. Before showing that (5) – (15) lead to the desired ancestral process, some intuition behind the role of each assumption is in order. (5) yields a limit process evolving in continuous time. Assumptions (6) – (8) ensure that the population sizes and time scales on demes are comparable. The conditions on the relative population sizes $\lambda_i(t)$ are sufficient to ensure finite waiting times between merger and migration events, and could be relaxed in specific examples. For exponential population growth, they hold as long as the growth rates on all demes coincide. For models in the domain of attraction of Kingman's coalescent, (7) will typically hold with $\gamma = 0$, while e.g. the Beta$(2-\alpha, \alpha)$-coalescents of Schweinsberg [2003] have $\gamma = 2 - \alpha$ (c.f. (4) and [Schweinsberg, 2003, Lemma 13]). The $\gamma \in [0, 1)$ condition ensures that (5) and (7) can hold simultaneously. Conditions (9) and (10) are well known to be necessary and sufficient for a Λ-coalescent limit, resulting in no more than four simultaneous multiple mergers in the diploid, biparental setting. (11) – (14) ensure that mutation, recombination, migration, and selection all take place on the coalescent time scale, while (15) disallows multiple selective branching events, as well as simultaneous selective and neutral merger events.

The aim is to show that the ancestry of a sample from the above particle system converges to a structured, time-inhomogeneous Ξ-Ancestral Influence Graph [Donnelly and Kurtz, 1999b] as $N \to \infty$, when time is measured in units of c_N. To establish this fact, we identify the limiting rates of coalescence, mutation, recombination, migration and branching due to selection, and show that these are the only dynamics which affect the ancestry of the process. Specifically, that $r \leq 4$ simultaneous mergers of sizes $b_1, \ldots, b_r$, with $2 \leq b_j \leq n_i$ at time t happen on deme i at rate

$$\frac{c_N \lambda_i(t)^{\gamma-1}}{\sum_{j=1}^D \lambda_j(0)^{\gamma+1}} \sum_{l=0}^{(n_i-b)\wedge(4-r)} \binom{n_i-b}{l} \frac{(4)_{r+l}}{4^{b+l}} \int_0^1 x^{b+l-2}(1-x)^{n_i-b-l} \Lambda_i(dx),$$

events in which one lineage branches into $4L+1$ lineages occur at rate $n_i\sigma_i/2$, branching into two lineages due to crossover recombination happens at rate $n_i\rho$, mutations occur ate rate $n_i\theta$, and that migration to deme $j \neq i$ happens at rate $n_i \frac{\lambda_j(t)}{\lambda_i(t)} m_{ji}$, where n_i is the number of lineages on deme i. Between migration events, the ancestries of subpopulations on different demes evolve independently. Convergence will then follow from a straightforward analogue of [Möhle and Sagitov, 2003, Theorem 4.2]. Throughout, we assume that our sample consists of n_i lineages on deme i, and that each lineage carries ancestral material on only one chromosome. This assumption is justified later by verifying that a separation of timescales phenomenon [Möhle, 1998] takes place, establishing that distinct chromosomes disperse to separate active lineages instantaneously on the coalescent time scale.

Multiple mergers via a single large family

By the Kingman formula for exchangeable, diploid offspring distributions [Möhle and Sagitov, 2003, equation (9)], the probability of $b \leq n_i$ chromosomes merging by belonging to the same family in the previous time step, and picking the same parental chromosome out of the four possibilities, is

$$4^{1-b} \frac{(M_N^{(i)}(\lfloor t/c_N \rfloor))_{n_i-b+1}}{(2M_N^{(i)}(\lfloor t/c_N \rfloor - 1))_{n_i}} \mathbb{E}[(\nu_1^{(i)})_b \nu_2^{(i)} \ldots \nu_{n_i-b+1}^{(i)}].$$

Analogously to [Möhle and Sagitov, 2003, equations (28) and (29)], conditions (9) and (10) imply that

$$\begin{aligned}\frac{M_N^{(i)}(\lfloor t/c_N \rfloor)^{n_i-b+1}}{2^{n_i-b+1} M_N^{(i)}(\lfloor t/c_N \rfloor - 1)^{n_i}} \mathbb{E}[(\nu_1^{(i)})_b \nu_2^{(i)} \ldots \nu_{n_i-b+1}^{(i)}] &= c_N^{(i)}(t) 4^{2-b} \int_0^1 x^{b-2}(1-x)^{n_i-b} \Lambda_i(dx) \\ &= \frac{c_N \lambda_i(t)^{\gamma-1} 4^{2-b}}{\sum_{j=1}^D \lambda_j(0)^{\gamma+1}} \int_0^1 x^{b-2}(1-x)^{n_i-b} \Lambda_i(dx),\end{aligned}$$

where the last step follows from (16). The rate of a particular combination of $r \leq 4$ simultaneous mergers with sizes $b_j \geq 2$ for $j \in [r]$ is obtained by summing over all ways in which such a merger can happen, resulting in the overall rate

$$\frac{c_N \lambda_i(t)^{\gamma-1}}{\sum_{j=1}^{D} \lambda_j(0)^{\gamma+1}} \sum_{l=0}^{(n_i-b)\wedge(4-r)} \binom{n_i-b}{l} \frac{(4)_{r+l}}{4^{b+l}} \int_0^1 x^{b+l-2}(1-x)^{n_i-b-l} \Lambda_i(dx) \tag{17}$$

where $b_1 + \ldots b_r = b \leq n_i$ [Birkner et al., 2013, equation (27)].

Multiple mergers via two or more large families

By (9), the probability of mergers via two or more large families, i.e. families with at least two offspring in the sample, is bounded from above by

$$\frac{1}{(2M_N^{(i)}(\lfloor t/c_N \rfloor - 1))_{n_i}} \sum_{j_1 \neq j_2 = 1}^{M_N^{(i)}(\lfloor t/c_N \rfloor)} \mathbb{E}\left[(\nu_{j_1}^{(i)} + \tilde{\beta}_{j_1}^{(i)})_2 (\nu_{j_2}^{(i)} + \tilde{\beta}_{j_2}^{(i)})_2 \left(\sum_{k=1}^{M_N^{(i)}(\lfloor t/c_N \rfloor)} \nu_k^{(i)} + \beta_k^{(i)} \right)^{n_i - 4} \right]$$
$$\leq \frac{(M_N^{(i)}(\lfloor t/c_N \rfloor))_2}{(2M_N^{(i)}(\lfloor t/c_N \rfloor - 1))_4} \mathbb{E}[(\nu_1^{(i)} + \tilde{\beta}_1^{(i)})_2 (\nu_j^{(i)} + \tilde{\beta}_2^{(i)})_2] = o(c_N).$$

A single migration event

The event that all n_i lineages belong to different families in the previous generation, and that one individual migrates from deme i to j in reverse time, has asymptotic probability

$$\left(1 - \sum_{k=1}^{D} m_{ik}^{(N)}\right)^{n_i - 1} n_i \frac{M_N^{(j)}(\lfloor t/c_N \rfloor - 1) m_{ji}^{(N)}}{M_N^{(i)}(\lfloor t/c_N \rfloor - 1)} \sim c_N n_i \frac{\lambda_j(t)}{\lambda_i(t)} m_{ji},$$

by (3), (8), and (13).

A similar calculation demonstrates that the analogous probability for more than one simultaneous migration event is $o(c_N)$, while combining the above with the first two calculations demonstrates that a single migration occurring simultaneously with one or more large families is also an $o(c_N)$ event.

A single mutation event

An analogous calculation to the migration case using (11) shows that the probability of one site mutating in the previous time step with no other accompanying events converges to $c_N n_i \theta$.

A single recombination event

Likewise, an analogous calculation to the migration case using (12) shows that the probability of one chromosome recombining in the previous time step with no other accompanying events converges to $c_N n_i \rho$.

A single branching event due to a selective birth

The probability of a lineage belonging to a selective birth by a family in the previous generation depends on the fitness of its parents, which is unknown. An elegant solution is to add selective events at the greatest possible rate, add the $4L$ chromosomes belonging to the two potential parents into the sample along with retaining the child lineage whenever a selection event happens, and track this extended sample to its *ultimate ancestor*: the most recent common ancestor of the original sample, as well as all potential selective parents encountered along the way [Krone and Neuhauser, 1997, Neuhauser and Krone, 1997]. The type of the ultimate ancestor can then be sampled from the stationary distribution of M (or any other desired initial law), with mutations occurring along lineages and alleles propagated to children as before. Now the alleles, and thus the fitness of the selective parents are known at each potential selective event, and the true ancestry of each child lineage can be assigned to either a randomly chosen parent with probability $\mathbb{E}[\beta_j^{(i)}]/\mathbb{E}[\tilde{\beta}_j^{(i)}]$, or to remain with the ongoing child lineage with the complementary probability.

From the point of view of the ancestral process, such selective branching events in which one lineage on deme $i \in [D]$ branches into $4L + 1$ lineages (corresponding to the single-chromosome child lineage, as well as the $4L$ parental chromosomes which immediately disperse into separate lineages due to separation of time scales) happens with asymptotic probability

$$\begin{aligned}
&\frac{1}{(2M_N^{(i)}(\lfloor t/c_N \rfloor - 1))_{n_i}} \sum_{\substack{j_1 \neq \ldots \neq j_{n_i} = 1 \\ \text{all distinct}}}^{M_N^{(i)}(\lfloor t/c_N \rfloor)} \mathbb{E}[\tilde{\beta}_{j_1}^{(i)} \nu_{j_2}^{(i)} \ldots \nu_{j_{n_i}}^{(i)}] \\
&= \frac{1}{(2M_N^{(i)}(\lfloor t/c_N \rfloor - 1))_{n_i}} \sum_{j=1}^{M_N^{(i)}(\lfloor t/c_N \rfloor)} \mathbb{E}\left[\tilde{\beta}_j^{(i)} \left(\sum_{k \neq j}^{M_N^{(i)}(\lfloor t/c_N \rfloor)} \nu_k^{(i)} \right)^{n_i - 1} \right] \\
&= \frac{M_N^{(i)}(\lfloor t/c_N \rfloor)}{(2M_N^{(i)}(\lfloor t/c_N \rfloor - 1))_{n_i}} \mathbb{E}\left[\tilde{\beta}_1^{(i)} \left(2M_N^{(i)}(\lfloor t/c_N \rfloor - 1) - \nu_1^{(i)} - \sum_{k=1}^{M_N^{(i)}(\lfloor t/c_N \rfloor)} \beta_k^{(i)} \right)^{n_i - 1} \right].
\end{aligned}$$

A binomial expansion followed by (14) and (15) yield

$$\begin{aligned}
&\frac{1}{(2M_N^{(i)}(\lfloor t/c_N \rfloor - 1))_{n_i}} \sum_{\substack{j_1 \neq \ldots \neq j_{n_i} = 1 \\ \text{all distinct}}}^{M_N^{(i)}(\lfloor t/c_N \rfloor)} \mathbb{E}[\tilde{\beta}_{j_1}^{(i)} \nu_{j_2}^{(i)} \ldots \nu_{j_{n_i}}^{(i)}] \\
&= \frac{M_N^{(i)}(\lfloor t/c_N \rfloor)}{(2M_N^{(i)}(\lfloor t/c_N \rfloor - 1))_{n_i}} \sum_{l=0}^{n_i - 1} \binom{n_i - 1}{l} [2M_N^{(i)}(\lfloor t/c_N \rfloor - 1)]^{n_i - 1 - l} (-1)^l \\
&\quad \times \mathbb{E}\left[\tilde{\beta}_1^{(i)} \left(\nu_1^{(i)} + \sum_{k=1}^{M_N^{(i)}(\lfloor t/c_N \rfloor)/2} \beta_k^{(i)} \right)^{l} \right] \\
&\sim \frac{M_N^{(i)}(\lfloor t/c_N \rfloor)}{2M_N^{(i)}(\lfloor t/c_N \rfloor - 1)} \mathbb{E}[\tilde{\beta}_1^{(i)}] + o(c_N) \sim c_N \frac{\sigma_i}{2} + o(c_N),
\end{aligned}$$

as required.

Multiple simultaneous branching events

Multiple simultaneous selective events can take place in one of three ways: two (or more) simultaneous selective births in the same family, two (or more) simultaneous selective births in a combination of families, or a combination of selective and neutral births in the same family. The probability of all three kinds of events is bounded above by

$$
\begin{aligned}
&\frac{M_N^{(i)}(\lfloor t/c_N\rfloor)}{(2M_N^{(i)}(\lfloor t/c_N\rfloor - 1))_{n_i}}\mathbb{E}\left[\tilde{\beta}_1^{(i)}\left\{2M_N^{(i)}(\lfloor t/c_N\rfloor)^{n_i-1} - \sum_{\substack{i_2\neq\ldots\neq i_a\neq 1\\ \text{all distinct}}}^{M_N^{(i)}(\lfloor t/c_N\rfloor)} \nu_{i_2}^{(i)}\ldots\nu_{i_a}^{(i)}\right\}\right]\\
&\leq \frac{M_N^{(i)}(\lfloor t/c_N\rfloor)}{(2M_N^{(i)}(\lfloor t/c_N\rfloor - 1))_{n_i}}\mathbb{E}\Bigg[\tilde{\beta}_1^{(i)}\Bigg\{2M_N^{(i)}(\lfloor t/c_N\rfloor - 1)^{n_i-1}\\
&\qquad\qquad\qquad - \left(2M_N^{(i)}(\lfloor t/c_N\rfloor - 1) - \sum_{j=1}^{M_N^{(i)}(\lfloor t/c_N\rfloor)}\tilde{\beta}_j^{(i)}\right)^{n_i-1}\Bigg\}\Bigg]\\
&= o(c_N),
\end{aligned}
$$

by a binomial expansion and (15).

Dispersal of chromosomes into distinct, single-marked individuals

Finally, we abandon the assumption that all lineages carry ancestral material on only one chromosome in order to verify the separation of time scales phenomenon. The probability that $n/2$ individuals with ancestral material both chromosomes (or so-called *double-marked* individuals) in a pair disperse into n parents, each of whom carries ancestral material on only one copy of the chromosome (so-called *single-marked* individuals), in the previous generation is $O(1)$. To see why, note that every individual is replaced at every time step, and individuals always inherit one chromosome from each parent. Thus, complete dispersal of $n/2$ double-marked individuals happens in one generation provided that all $n/2$ individuals originate from different families, which has probability at least

$$\prod_{i=1}^{D}\frac{(M_N^{(i)}(t))_{n_i}}{(2M_N^{(i)}(t-1))_{n_i}}\mathbb{E}[\nu_1^{(i)}\ldots\nu_{n_i}^{(i)}] = O(1).$$

Likewise, the probability of two active chromosomes splitting apart into distinct ancestors is $1/2 = O(1)$ because assignments of parents to chromosomes is done independently and uniformly at random. Hence, the probability of a lineage with $2L$ ancestral chromosomes dispersing into $2L$ lineages with a single ancestral chromosome each in at most $2L-1$ generations happens with probability at least

$$\left(\frac{1}{2}\right)^{2L}\prod_{t=1}^{2L}\prod_{i=1}^{D}\frac{(M_N^{(i)}(t))_{n_i}}{(2M_N^{(i)}(t-1))_{n_i}}\mathbb{E}[\nu_1^{(i)}\ldots\nu_{n_i}^{(i)}] = O(1).$$

Probabilities of merger, recombination, selection or migration events were all established above to be $O(c_N)$, and thus the probability of complete dispersal before any merger, recombination, selection or migration events is of order

$$\frac{1}{1+Ac_N}\to 1,$$

where $A > 0$ is a constant independent of both n and N. Thus an analogue of the separation of timescales result in [Möhle, 1998] holds, which justifies considering only single-marked configurations in the previous computations of transition probabilities.

4 Robustness results

The following three subsections quantify the respective effect of selection, recombination, and population structure on the sampling distribution of the singleton-tail statistic. Each subsection specialises the model of Section 3 to consist of only the relevant force by a particular choice of parameters. We assume the model of Schweinsberg [2003] for the evolution of the population, and thus consider a one-dimensional family of coalescents specified by $\Lambda_i(dx) = \text{Beta}(2 - \alpha, \alpha)(dx)$ in (10) for $\alpha \in (1, 2)$, with corresponding time scaling $c_N \sim N^{1-\alpha}$ and $\gamma = 2 - \alpha$ in (7). Under the alternative hypothesis $\alpha < 2$, the population sizes on demes will be constant, i.e. $\lambda_i(t) = d_i$ for relative deme sizes $d_1 + \ldots + d_D = 1$. Under the null hypothesis $\alpha = 2$, populations on demes will undergo exponential growth forwards in time, corresponding to $M_N^{(i)}(t) := \lfloor N d_i (1 + \gamma_N)^{-t} \rfloor$, resulting in $\lambda_i(t) = d_i e^{-\gamma t}$ for the population-rescaled growth rate $\gamma = \lim_{N \to \infty} \gamma_N / c_N$.

It will also be necessary to distinguish between two kinds of data sets: simulated data used to fit KDEs to approximate likelihoods, and compute the quantile q_ω in (1), as well as observed data, which will also be simulated in this instance, but which will typically be a biological data set. We will refer to the former as calibration data, and the latter as pseudo-observed data. Pseudo-observed data is reserved solely for plugging into KDE approximations of likelihoods (computed from calibration data) to obtain likelihood ratio test statistics. A C++ implementation of the algorithm used to generate the data in this section is available at https://github.com/JereKoskela/Beta-Xi-Sim.

We set the number of simulation replicates per model at 1000 (note that Θ_0 contains 133 models, and Θ_1 a further 41), the sample size at $n = 500$, the lumped tail cutoff at $k = 15$, and assume the true mutation rate is known. The number of unlinked chromosomes per sample is set to 23 to match the number of linkage groups in Atlantic cod [Tørresen et al., 2017, Supplementary Table 3] — an organism for which multiple merger have frequently been suggested as an important evolutionary mechanism [Steinrücken et al., 2013, Tellier and Lemaire, 2014]. Results are averaged across chromosomes as outlined in Section 2. The lengths of the 23 chromosomes have also been set (by multiplying the total rate of mutation on each chromosome by the number of sites it contains) to match those reported in [Tørresen et al., 2017, Supplementary Table 3]. The approximate size of hypothesis tests is set at $\omega = 0.01$ throughout.

4.1 Weak selection

In this section we consider the model of Section 3 with a single deme ($D = 1$), and no recombination ($\rho = 0$). The resulting process is a Ξ-coalescent analogue of the Complex Selection Graph (CSG) Fearnhead [2003]. We compute realisations of the singleton-tail statistic by assuming that neutral, infinitely-many-sites mutations occur on each chromosome along the branches of the realised non-neutral tree sampled from the Ξ-CSG, but that the selective types of individuals are unobserved. This assumption is reasonable if the fitness of individuals cannot be observed, or if mutations with a fitness effect are either much less frequent than neutral mutations, or occur in unobserved regions.

Figure 1 shows the sampling distributions of the neutral and non-neutral models. The fitness model assumes two alleles, a and A, with each chromosome pair contributing fitness $\sigma > 0$ if either parent carries at least one A allele at that pair, and 0 otherwise. The selection rates are necessarily low, because the cost of simulating the ASGs is known to increase exponentially in σ [Fearnhead, 2001, Appendix A]. Efficiency gains resulting from perfect simulation techniques [Fearnhead, 2001, 2003] cannot be employed because they rely on terminating the simulation before reaching the MRCA, and thus the SFS cannot be resolved.

The results in Figure 1 show striking agreement between sampling distributions in the neutral and selective cases. We also conducted the hypothesis test (1) using calibration data simulated from a neutral model, and applied the resulting misspecified test to pseudo-observed data simulated from a model with weak selection. Figure 2 shows that the performance of the test was excellent, with high power and size well below the formal threshold of $\omega = 0.01$ for the majority

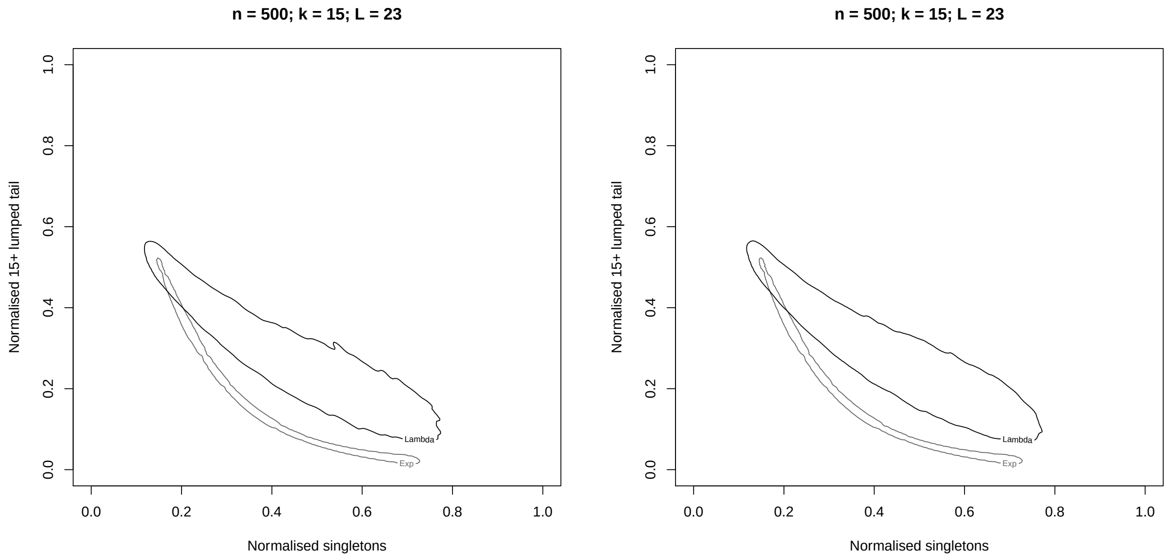

Figure 1: 99^{th} percentiles of KDEs fitted to 1000 realisations of the singleton-tail statistic for each model in Θ_0 and Θ_1. Each sample consists of 23 chromosomes, and (Left) $\sigma = 0$, or (Right) $\sigma \in (0.0000016, 0.0008)$ per chromosome pair as α varies from 1 to 2.

of the parameter ranges.

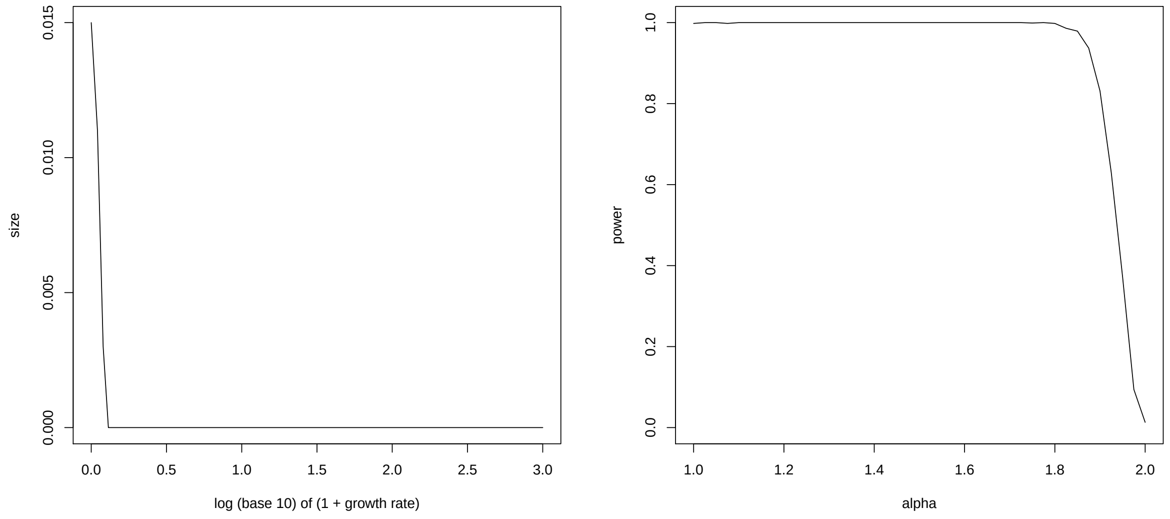

Figure 2: Empirical size (Left) and power (Right) of a Θ_0 vs Θ_1 test conducted using calibration data simulated under neutral models, but applied to pseudo-observed data simulated from selective models. The simulation parameters are as in Figure 1.

To investigate the effect of a larger selection coefficient, we also simulated realisations of the singleton-tail statistic under a single chromosome model. In this setting, each selective branching event results in five lineages, as opposed to $4 \times 23 + 1 = 93$ as in the 23 chromosome case. The sampling variance under a single locus is too large for a powerful statistical test, but Figure 3 demonstrates that the sampling distributions with and without selection remain very similar. Taken together, these simulations show that the distribution of realised relative branch lengths under the CSG is similar to that under a neutral coalescent, at least for external branches, as well as for the oldest branches before the MRCA is reached. Hence, the singleton-tail statistic cannot be used to detect weak selection, but can discriminate between population growth and Ξ-coalescents without knowing whether weak selection is taking place.

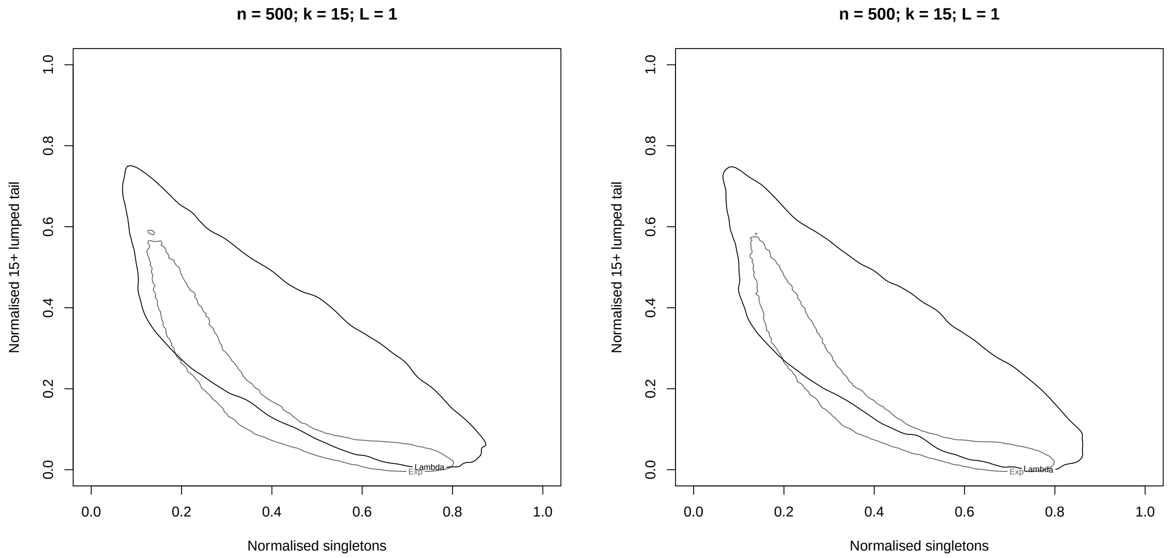

Figure 3: 99^{th} percentiles of KDEs fitted to 1000 realisations of the singleton-tail statistic for each model in Θ_0 and Θ_1. Each sample consists of one chromosome, and (Left) $\sigma = 0$, or (Right) $\sigma \in (0.0024, 1.2)$ as α varies from 1 to 2.

4.2 Recombination

In this section we consider the model of Section 3 with a single deme ($D = 1$), and no selection ($\sigma_i \equiv 0$). Realisations of the singleton-tail statistic are computed by assuming a neutral, infinitely-many-sites mutation model along the branches of the realised Ξ-Ancestral Recombination Graph.

Figure 4 presents a comparison between models with and without recombination. As was the case with weak selection (see Figure 1), the presence of recombination makes no discernible difference to the sampling distribution of the singleton-tail statistic, although the distribution of intermediate SFS entries was observed to be different (results not shown). Figure 5 demonstrates that the size and power of statistical tests are unaffected when a misspecified model which wrongly neglects recombination is used to generate calibration data, and the hypothesis test is conducted on pseudo-observed data with recombination.

4.3 Population structure

In this section we consider the model of Section 3 with a no selection ($\sigma_i \equiv 0$), no recombination ($\rho = 0$), and two different patterns of population structure.

Figure 6 shows sampling distributions corresponding to a four deme model with symmetric migration between all pairs of demes, as well as a two deme model with asymmetric migration. The contours differ markedly from the panmictic results in Figures 1 and 4, and also from each other. Figure 7 demonstrates that misspecifying spatial structure results in very poor performance of the hypothesis test, with both the size and power curves showing complex behaviour

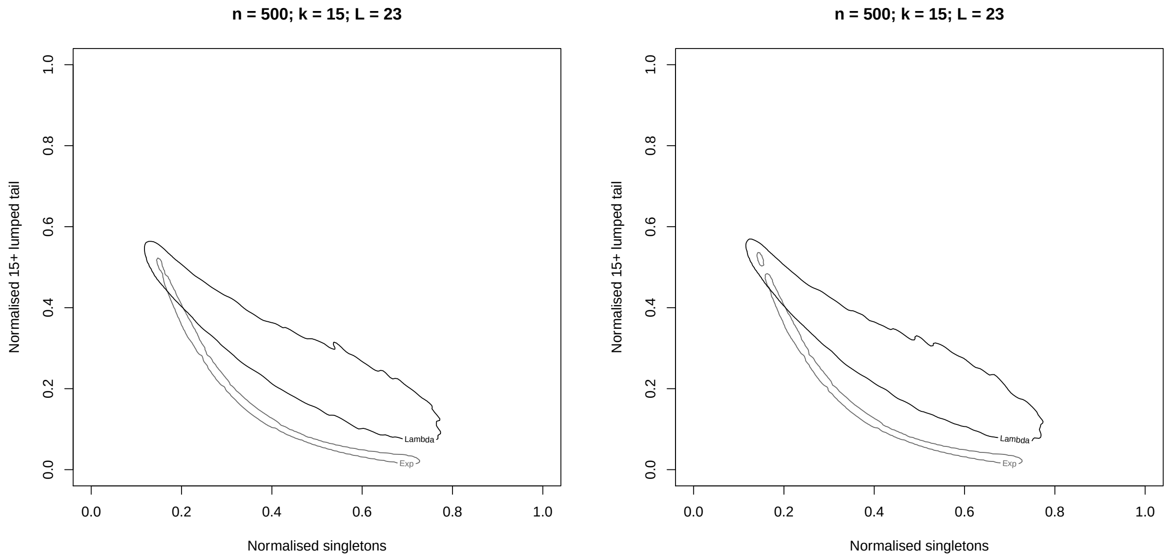

Figure 4: 99^{th} percentiles of KDEs fitted to 1000 realisations of the singleton-tail statistic for each model in Θ_0 and Θ_1. (Left) $\rho = 0$. (Right) $\rho \in (0.001, 0.5)$ as α varies from 1 to 2.

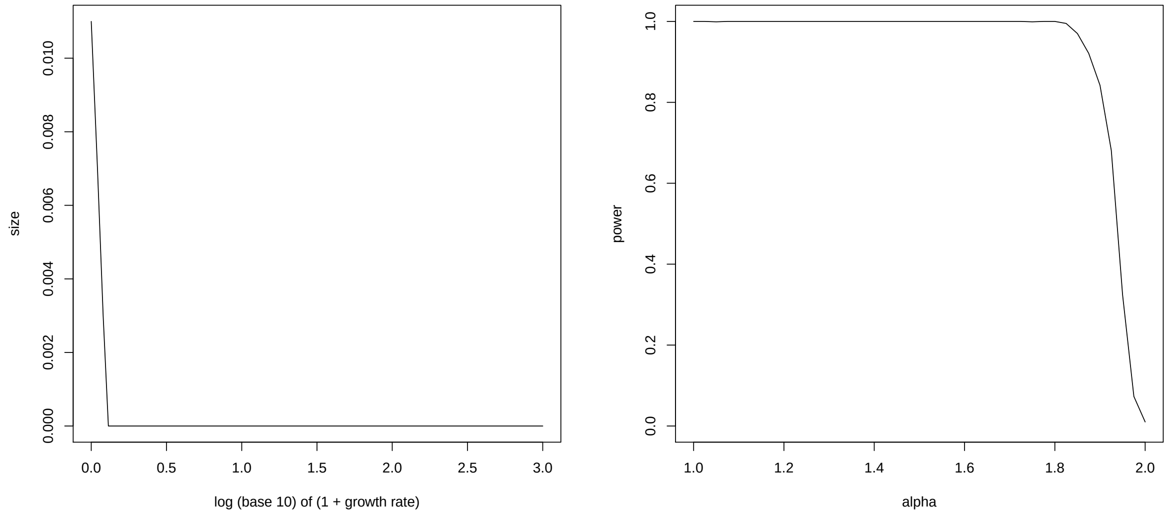

Figure 5: Empirical size (Left) and power (Right) of a Θ_0 vs Θ_1 test conducted using calibration data from models without recombination, but applied to pseudo-observed data from models with recombination. The simulation parameters are as in Figure 4.

that depends on the patterns of overlap between the distribution of the misspecified calibration data, and the pseudo-observed data.

5 Distinguishing high fecundity from selective sweeps

This section focuses on distinguishing multiple mergers due to selective sweeps from multiple mergers due to high fecundity. The high fecundity model is the Ξ-coalescent introduced in Section 3 with a single deme ($D = 1$), no recombination $\rho = 0$, no selection $\sigma_i = 0$, and a constant population size $\lambda_1(t) = 1$. For selective sweeps, we assume a population of constant

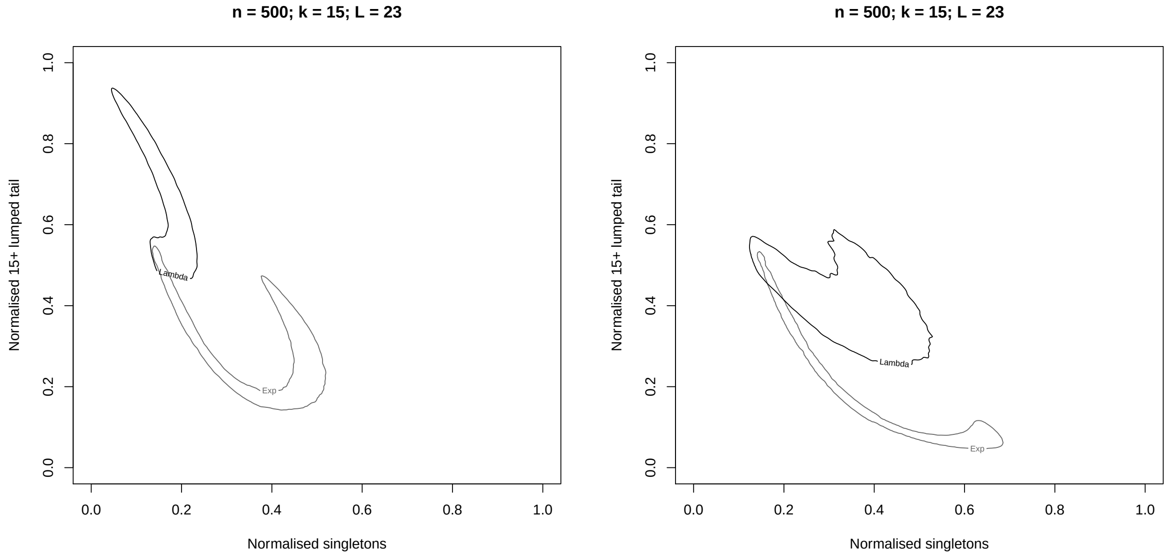

Figure 6: 99^{th} percentiles of KDEs fitted to 1000 realisations of the singleton-tail statistic for each model in Θ_0 and Θ_1. (Left) Four demes with equal population sizes and symmetric migration between all pairs of demes at reverse-time rate $\tilde{m} \in (0.01, 5)$ as α varies from 1 to 2. (Right) Two demes with relative population sizes $(0.75, 0.25)$, and reverse-time migration rates ranging from $(\tilde{m}_{12}, \tilde{m}_{21}) = (0.01, 0.03)$ when $\alpha = 1$ to $(\tilde{m}_{12}, \tilde{m}_{21}) = (5, 15)$ when $\alpha = 2$.

size evolving in non-overlapping generations, in which mutations providing a selective advantage $x \in (0, 1)$ occur at points of a Poisson process with rate $x^{-2}\,\mathrm{Beta}(2 - \alpha, \alpha)(dx)$, and sweep to fixation instantaneously on the coalescent time scale.

We also assume that recombination within chromosomes results in incomplete sweeps, so that when viewed backwards in time each individual has a random chance to participate in the merger resulting from each sweep. Recombination is specified implicitly by setting the probability of a lineage participating in a sweep arising from a mutation with advantage $x \in (0, 1)$ to x. Genetic material that is unlinked to the beneficial mutation escapes the selective sweep, and thus multiple mergers affect one chromosome at a time. Neutral mutations continue to accrue along ancestral branches according to the infinite sites model with mutation rate $\theta > 0$. When the population is diploid and biparental, these dynamics result in an ancestral process in which the marginal coalescent at each locus is the Ξ-coalescent with merger rates given by (17).

Remark 4. The model described above has not been derived as a scaling limit of a finite population model of evolution. Instead, it has been chosen to make the task of distinguishing between selective sweeps and high fecundity as difficult as possible. For the same reason, we also scale the mutation rate as $\theta \propto \lim_{N \to \infty} N^{\alpha - 1} \mu_N$ as in the model of Schweinsberg [2003]. For biological motivation, note that this model closely resembles the Λ-coalescent of Durrett and Schweinsberg [2005], which was derived as a scaling limit of finite population models undergoing selective sweeps and recombination in much the same way as above. However, their convergence result can only be used to obtain Λ-coalescents in which Λ has an atom at 0, and hence it cannot be immediately used to obtain our model [Durrett and Schweinsberg, 2005, Example 2.5]. A model akin to ours could be obtained as a similar scaling limit by letting selective sweeps occur more frequently than the time scale of pairwise coalescence, thus causing the atom at 0 to vanish in the large population limit [Gillespie, 2000, Durrett and Schweinsberg, 2005].

We fix our null hypothesis as the class of selective sweep models described above with the parameter $\alpha \in (1, 2)$ discretised as in (2). The alternative hypothesis is the high fecundity

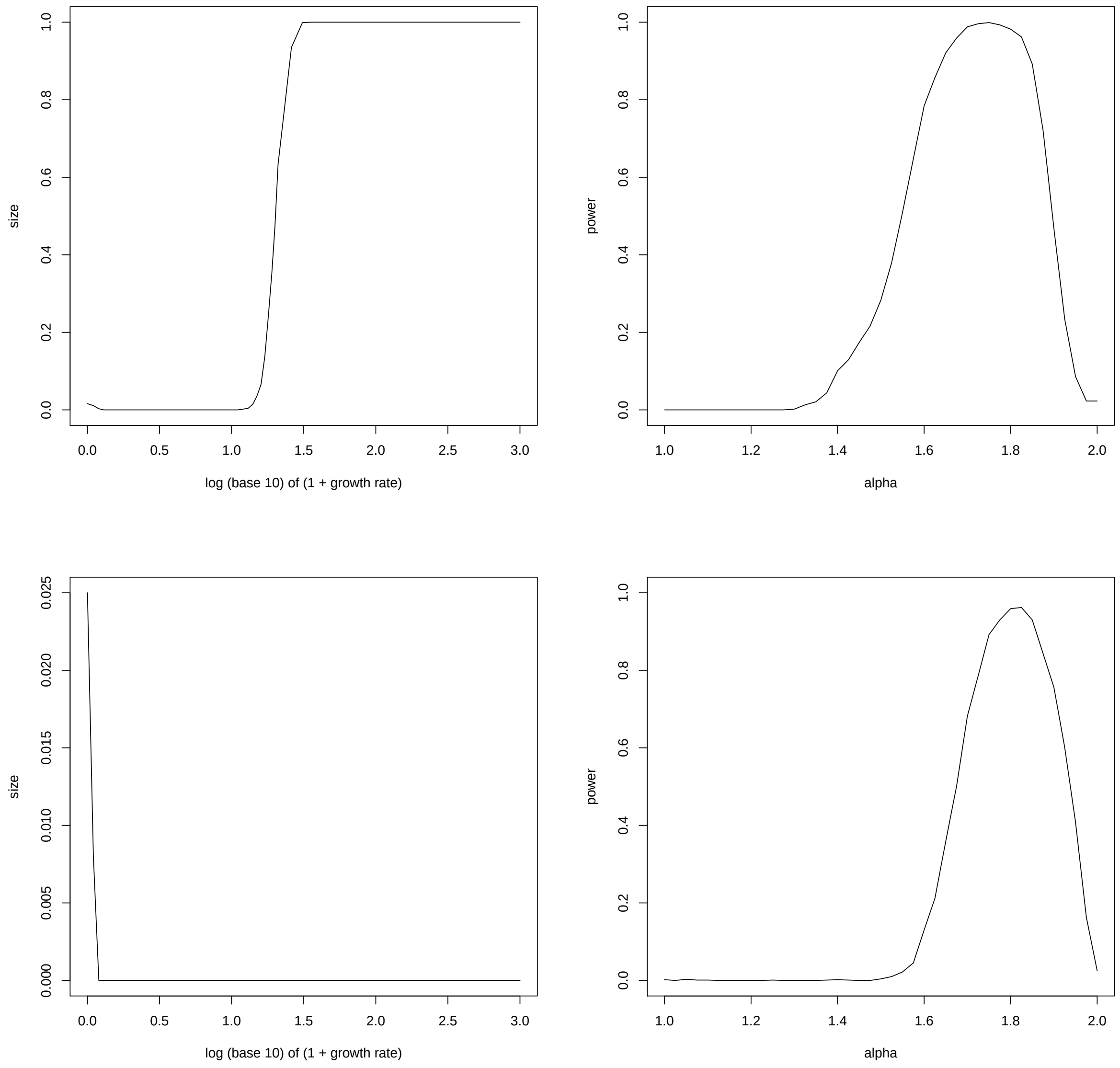

Figure 7: (Top Row) Empirical size (Left) and power (Right) of a test of Θ_0 vs Θ_1 conducted using calibration data from panmictic models, but applied to pseudo-observed data from four deme models. (Bottom Row) Empirical size (Left) and power (Right) of a test of Θ_0 vs Θ_1 conducted using calibration data from four deme models, but applied to pseudo-observed data from two deme models. The simulation parameters in both cases are as in Figure 6.

Ξ-coalescent described at the beginning of this section. The only difference between the two model classes is whether coalescence times at unlinked chromosomes are independent (under the selective sweep model), or positively correlated (under the high fecundity model). The marginal coalescents at each chromosome coincide.

Figure 8 demonstrates that the singleton-tail statistic exhibits higher sampling variance under the alternative hypothesis than under the null due to positive correlation of coalescence times between unlinked chromosomes. While the sampling distributions under both hypotheses centre on the same mean, increased variance means that the null hypothesis can be correctly rejected with moderate power of around 30% for the majority of the parameter range. Reversing the roles of the hypotheses caused the power to vanish, as the bulk of the sampling distribution under the alternative (selective sweep) hypothesis is fully contained in that of the null (high

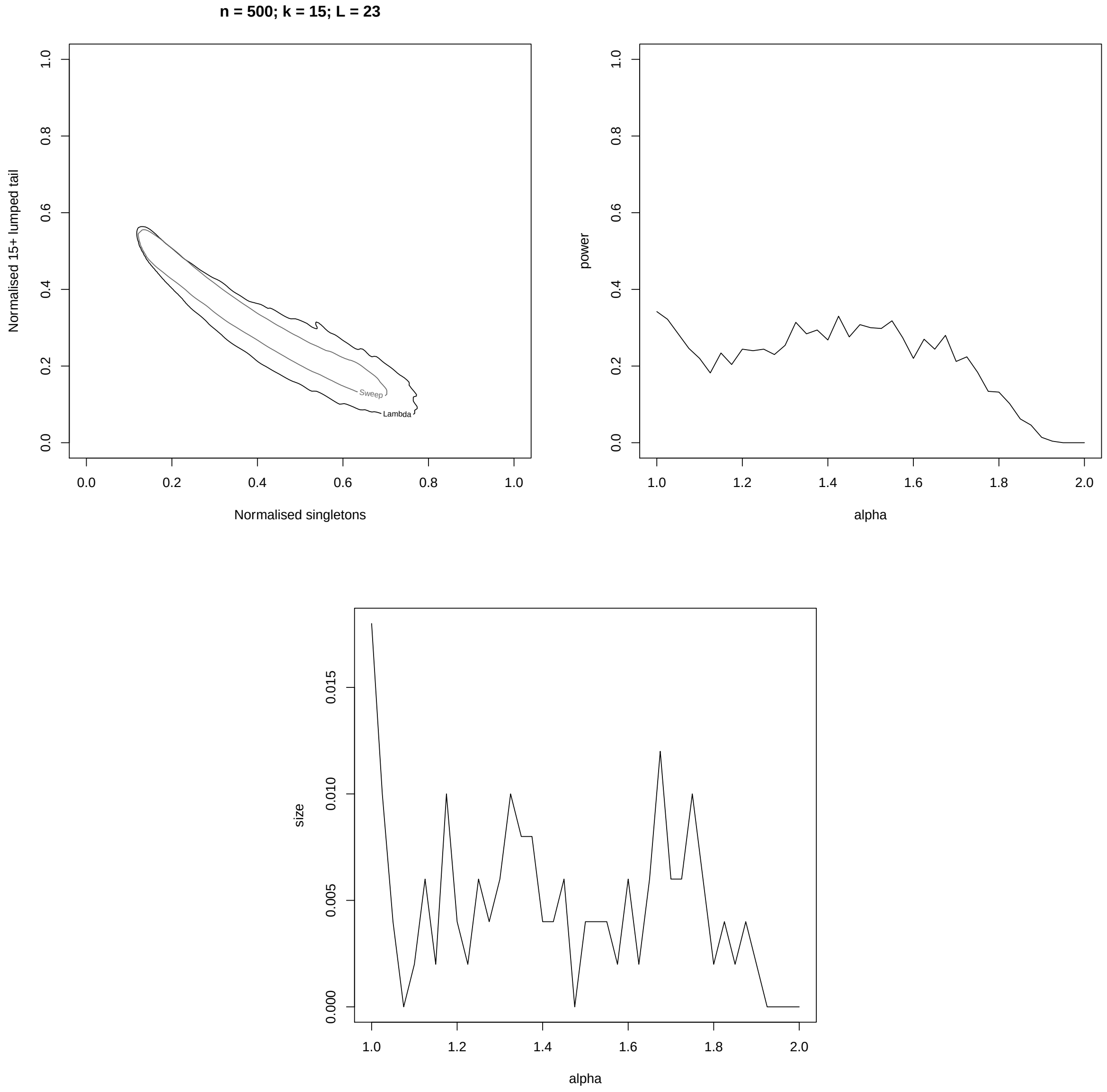

Figure 8: (Top Left) 99^{th} percentile of a kernel density estimator fitted to 1000 realisations of the singleton-tail statistic under each model in Θ_0 and Θ_1. (Top Right) Empirical power and (Bottom) empirical size of the likelihood ratio test (1).

fecundity) hypothesis (results not shown).

6 Discussion

We have derived a coalescent model of population growth and high fecundity involving multiple chromosomes under the joint effect of weak selection, recombination, and spatial structure. We studied the effect of these three confounders on the ability of the singleton-tail statistic Koskela [2018] to distinguish between population growth models and Ξ-coalescent models of high fecundity.

Crossover recombination and weak natural selection had no visible effect on the sampling distribution of the singleton-tail statistic. Therefore, the statistic retained its ability to distinguish high fecundity from population growth with high power based on multi-locus data under these confounders. Moreover, model selection can be based on calibration data simulated from a neutral model without recombination. This reduces the number of nuisance parameters, and yields significant efficiency gains because selection and recombination are very expensive to simulate. The computational speedup compensates for the relatively large sample size of 500, which appears to be necessary for a high-powered test, as samples of size 100 were shown in Koskela [2018] to have noticeably lower power even without any of the confounders considered in this article.

Population structure had a significant effect on the sampling distribution of the singleton-tail statistic. Misspecifying population structure when simulating calibration data rendered the hypothesis test inaccurate, with erratic false positives and false negative probabilities. It is well known that spatial structure results in an excess of intermediate and high frequency polymorphisms under the Kingman coalescent [Wakeley and Alicar, 2001, De and Durrett, 2007]. Our results confirm that similar phenomena also hold for Ξ-coalescents (Figure 6 shows a clear excess of high frequency polymorphisms and deficit of singletons), and that the exact amount of excess is sensitive to the details of the population structure. This finding motivates research into methods which can infer population structure without assuming a particular coalescent or growth scenario.

Finally, we investigated the ability of the singleton-tail statistic to distinguish multiple mergers due to selective sweeps from multiple mergers due to high fecundity. This is a challenging problem because the marginal models at single chromosomes coincide under the two hypotheses. However, ancestral trees at unlinked chromosomes are independent under selective sweeps, which affect the genome locally, and positively correlated under high fecundity which affects all loci simultaneously (c.f. [Koskela, 2018, Remark 1] for a formal justification). Positive correlation increases the sampling variance of the multi-locus singleton-tail statistic, which enabled us to distinguish a high fecundity alternative hypothesis from a selective sweep null hypothesis with moderate power. Reversing the roles of the hypotheses caused the power of the test to vanish, and thus model selection can only be successfully performed in one direction based on this method.

Acknowledgements

The authors are grateful to Jochen Blath and Bjarki Eldon for discussions concerning generalised Beta-Ξ-coalescent models, and to Fabian Freund for pointing out an error in an earlier version of the manuscript. JK was supported by Deutsche Forschungsgemeinschaft (DFG) grant BL 1105/3-2 as part of SPP Priority Programme 1590, and by EPSRC grant EP/R044732/1. MWB was supported by DFG grant BL 1105/5-1 as part of SPP Priority Programme 1590. Both authors were also supported by DFG SPP Priority Programme 1819 start-up module grant "Population genomics of highly fecund codfish".

References

E Árnason. Mitochondrial cytochrome *b* variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. *Genetics*, 166:1871–1885, 2004.

E Baake, U Lenz, and A Wakolbinger. The common ancestor type distribution of a Λ-Wright-Fisher process with selection and mutation. *Electron Commun Probab*, 21(59):1–16, 2016.

A T Beckenbach. Mitochondrial haplotype frequencies in oysters: neutral alternatives to selection models. In B Golding, editor, *Non-Neutral Evolution*, pages 188–198. Chapman & Hall, New York, 1994.

M Birkner, J Blath, and M Steinrücken. Importance sampling for Lambda-coalescents in the infinitely many sites model. *Theor Pop Biol*, 79:155–173, 2011.

M Birkner, J Blath, and B Eldon. An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, 193:255–290, 2013.

A De and R Durrett. Stepping-stone spatial structure causes slow decay of linkage disequilibrium and shifts the site frequency spectrum. *Genetics*, 176:969–981, 2007.

P Donnelly and T G Kurtz. Particle representations for measure-valued population models. *Ann Probab*, 27:166–205, 1999a.

P Donnelly and T G Kurtz. Genealogical processes for Fleming-Viot models with selection and recombination. *Ann Appl Probab*, 9:1091–1148, 1999b.

T Duong and M L Hazelton. Plug-in bandwidth matrices for bivariate kernel density estimation. *J. Nonparametr. Statist.*, 15:17–30, 2003.

R Durrett and J Schweinsberg. A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stoch Proc Appl*, 115:1628–1657, 2005.

B Eldon. Structured coalescent processes from a modified Moran model with large offspring numbers. *Theor Pop Biol*, 76:92–104, 2009.

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

B Eldon, M Birkner, J Blath, and F Freund. Can the site frequency spectrum distinguish exponential population growth from multiple-merger coalescents. *Genetics*, 199(3):841–856, 2015.

P Fearnhead. Perfect simulation from population genetic models with selection. *Theor Pop Biol*, 59:263–279, 2001.

P Fearnhead. Ancestral process for non-neutral models of complex diseases. *Theor Pop Biol*, 63:115–130, 2003.

Y X Fu. Statistical properties of segregating sites. *Theor Pop Biol*, 48:172–197, 1995.

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

R C Griffiths and P Marjoram. An ancestral recombination graph. In P Donnelly and S Tavaré, editors, *Progress in population genetics and human evolution*, pages 257–270. Springer Verlag, Berlin, 1997.

D Hedgecock and A I Pudovkin. Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. *Bull Mar Sci*, 87:971–1002, 2011.

H M Herbots. The structured coalescent. In P Donnelly and S Tavaré, editors, *Progress in population genetics*, pages 231–255. Springer, New York, 1997.

R R Hudson. Properties of a neutral allele model with intragenic recombination. *Theor Pop Biol*, 23:183–201, 1983a.

R R Hudson. Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, 37:203–217, 1983b.

J F C Kingman. The coalescent. *Stoch Proc Appl*, 13:235–248, 1982a.

J F C Kingman. Exchangeability and the evolution of large populations. In G Koch and F Spizzichino, editors, *Exchangeability in probability and statistics*, pages 97–112. North-Holland, Amsterdam, 1982b.

J F C Kingman. On the genealogy of large populations. *J Appl Probab*, 19A:27–43, 1982c.

J Koskela. Multi-locus data distinguishes between population growth and multiple merger coalescents. *Stat Appl Genet Mol Biol*, 17(3):20170011, 2018.

S M Krone and C Neuhauser. Ancestral processes with selection. *Theor Pop Biol*, 51:210–237, 1997.

V Limic and A Sturm. The spatial Λ-coalescent. *Electron. J. Probab.*, 11:363–393, 2006.

S Matuszewski, M E Hildebrandt, G Achaz, and J D Jensen. Coalescent processes with skewed offspring distributions and non-equilibrium demography. *Genetics*, 208(1):1323–1338, 2018.

M Möhle. A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Adv in Appl Probab*, 30:493–512, 1998.

M Möhle. The coalescent in population models with time-inhomogeneous environment. *Stochastic Process Appl*, 97:199–227, 2002.

M Möhle and S Sagitov. Classification of coalescent processes for haploid exchangeable coalescent processes. *Ann Probab*, 29:1547–1562, 2001.

M Möhle and S Sagitov. Coalescent patterns in diploid exchangeable population models. *J Math Biol*, 47:337–352, 2003.

C Neuhauser and S M Krone. The genealogy of samples in models with selection. *Genetics*, 145:519–534, 1997.

J Pitman. Coalescents with multiple collisions. *Ann Probab*, 27:1870–1902, 1999.

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

O Sargsyan and J Wakeley. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor Pop Biol*, 74:104–114, 2008.

J Schweinsberg. Coalescents with simultaneous multiple collisions. *Electron J Prob*, 5:1–50, 2000.

J Schweinsberg. Coalescent processes obtained from supercritical Galton-Watson processes. *Stoch Proc Appl*, 106:107–139, 2003.

M Steinrücken, M Birkner, and J Blath. Analysis of DNA sequence variation within marine species using beta-coalescents. *Theor Pop Biol*, 87:15–24, 2013.

F Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105: 437–460, 1983.

A Tellier and C Lemaire. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Mol Ecol*, 23:2637–2652, 2014.

O K Tørresen, B Star, S Jentoft, W B Reinar, H Grove, J R Miller, B P Walenz, J Knight, J M Ekholm, P Peluso, R B Edvardsen, A Tooming-Klunderud, M Skage, S Lien, K S Jakobsen, and A J Nederbragt. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics*, 18(1):95, 2017.

J Wakeley and N Alicar. Gene genealogies in a metapopulation. *Genetics*, 159:893–905, 2001.

G A Watterson. On the number of segregating sites in genetical models without recombination. *Theor Pop Biol*, 7:1539–1546, 1975.