Statistical Properties of the Site-Frequency Spectrum Associated with Λ -Coalescents

Matthias Birkner,* Jochen Blath,[†] and Bjarki Eldon^{*,1}

*Institut für Mathematik, Johannes-Gutenberg-Universität, 55099 Mainz, Germany, and [†]Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany

ABSTRACT Statistical properties of the site-frequency spectrum associated with Λ -coalescents are our objects of study. In particular, we derive recursions for the expected value, variance, and covariance of the spectrum, extending earlier results of Fu (1995) for the classical Kingman coalescent. Estimating coalescent parameters introduced by certain Λ -coalescents for data sets too large for full-likelihood methods is our focus. The recursions for the expected values we obtain can be used to find the parameter values that give the best fit to the observed frequency spectrum. The expected values are also used to approximate the probability a (derived) mutation arises on a branch subtending a given number of leaves (DNA sequences), allowing us to apply a pseudolikelihood inference to estimate coalescence parameters associated with certain subclasses of Λ -coalescents. The properties of the pseudolikelihood approach are investigated on simulated as well as real mtDNA data sets for the high-fecundity Atlantic cod (*Gadus morhua*). Our results for two subclasses of Λ -coalescents show that one can distinguish these subclasses from the Kingman coalescent, as well as between the Λ -subclasses, even for a moderate (maybe a few hundred) sample size.

LARGE offspring number population models have recently been proposed as appropriate models with which to investigate high-fecundity natural populations. Some marine populations may belong to the class of high-fecundity populations, including Pacific oysters (*Crassostrea gigas*) (Beckenbach 1994; Li and Hedgecock 1998; Boudry *et al.* 2002), white sea bream (*Diplodus sargus*) (Planes and Lenfant 2002), and Atlantic cod (*Gadus morhua*) (Árnason 2004). Oysters feature in Williams (1975)'s elm and oyster model as an example of a high-fecundity population. Indeed, high-fecundity populations are discussed at length by Williams (1975) when comparing the benefits of sexual vs. asexual reproduction. Avise *et al.* (1988) compare genetic distances for mitochondrial (mt)DNA variation for three vertebrate species, american eels (*Anguilla rostrata*), hardhead catfish (*Arius felis*), and red-winged blackbirds (*Agelaius phoeniceus*), and conclude that historical effective population sizes may have been much lower than current census size. Low effective population size compared to census population size observed

for certain marine populations in particular (e.g., Hedgecock *et al.* 1992), and reviewed by Hedgecock and Pudovkin (2011), may be evidence of high variance in offspring distribution. Indeed, Hedrick (2005) observes that low effective population size results from high variance in reproductive success in a population with large census size. High fecundity may also be a way for certain marine organisms with broadcast spawning to compensate for high mortality rate among juveniles, thus exhibiting type III survivorship curves.

Multiple-merger coalescent processes, so-called Λ - and Ξ -coalescents, arise naturally from large offspring number models (Donnelly and Kurtz 1999; Sagitov 1999, 2003; Möhle and Sagitov 2001; Eldon and Wakeley 2006; Sargsyan and Wakeley 2008; Birkner and Blath 2009; Huillet and Möhle 2011a). A key assumption to obtain a multiple-merger coalescent process from a large offspring number model is that the number of offspring of an individual can be up to the order of the population size with nonnegligible probability. In contrast to the classical Kingman coalescent framework, which has served as the standard null model in population genetics since its introduction (Kingman 1982), Λ -coalescents allow multiple instead of only binary collisions of ancestral lineages, in which the group of lineages coalescing in a given instance may include anywhere from two to all of the current active lineages. Selective sweeps

Copyright © 2013 by the Genetics Society of America
doi: 10.1534/genetics.113.156612

Manuscript received May 26, 2013; accepted for publication August 28, 2013

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.156612/-DC1>.

¹Corresponding author: Institut für Mathematik, Technische Universität Berlin, Strasse des 17. Juni 136, 10623 Berlin, Germany. E-mail: eldon@math.tu-berlin.de

(Durrett and Schweinsberg 2004, 2005; Coop and Ralph 2012), or large-scale spatial extinction and recolonization events (Barton *et al.* 2010, 2013), can also give rise to multiple-merger coalescent processes. In some cases, one naturally encounters the even more general Ξ -coalescents (Schweinsberg 2000), allowing *simultaneous* multiple mergers at a time, *i.e.*, when distinct groups of ancestral lineages coalesce at the same time to distinct ancestors. Models of severe bottlenecks (Birkner *et al.* 2009) or diploidy (Möhle and Sagitov 2003) and recombination in conjunction with distinct loci can give rise to Ξ -coalescents. By way of example, a Ξ -coalescent admitting up to quadrifold simultaneous mergers is obtained from a diploid multilocus model with a skewed (high variance) offspring distribution (Birkner *et al.* 2013). Möhle and Sagitov (2001) provide a full mathematical classification of coalescent limits (as population size tends to infinity) of Cannings' exchangeable population models.

The statistical methodology to test for the presence and particularities of the effects of multiple mergers, or to distinguish between different underlying Λ - and Ξ -coalescents within a family of models, is, however, still largely missing. Hedgecock and Pudovkin (2011) argue that large offspring number models are the right framework in which to study high-fecundity marine populations. Some large offspring number models (*e.g.*, Sagitov 2003; Eldon and Wakeley 2006) introduce new parameters (*coalescence* parameters) determining the size and intensity of large offspring number events, *i.e.*, when individuals have very many offspring. Estimating the coalescence parameters from genetic data may give us a way to distinguish between different population models. Indeed, Hedgecock and Pudovkin (2011) call for the development of such inference methods, in particular to infer parameters, to assess the quality of the estimators, and to construct hypothesis tests and confidence bounds.

On one hand are summary-statistics-based estimators for the parameter of certain classes of Λ -coalescents (Eldon 2011) in the infinite-sites model, such as the number of segregating sites, which may not represent a sufficient statistic, but are computationally efficient, and could be used to obtain prior information. On the other hand, there are full-likelihood-based point estimators, derived via relatively complex recursions—which are, in principle, amenable to Monte Carlo methods—taking all available information into account, but whose practical treatment becomes computationally prohibitive even for medium-sized data sets (Birkner and Blath 2008; Birkner *et al.* 2011). A small empirical simulation study for the latter can be found in Steinrücken (2009). For both types of estimators, an analysis of their statistical properties in the Λ case is still largely missing.

A summary statistic that contains much of the essential (in particular, topological) information of the underlying genealogical tree of the data is the *site-frequency spectrum* (SFS) (however, one may want to consult, *e.g.*, Myers *et al.* 2008 for theoretical limitations). The spectrum is a key quantity when applying coalescent theory in inference; see, *e.g.*, Wakeley (2007) for an overview and a discussion of the relation of

the SFS with various other (simpler) summary statistics. The statistical properties of the SFS under the Kingman coalescent have been investigated in several studies (Fu 1995; Griffiths and Tavaré 1998; Živković and Wiehe 2008; Achaz 2009; Sainudiin *et al.* 2011). Extending in particular the results of Fu (1995) to Λ -coalescents is our aim. Recursions for the expected value, the variance, and the covariance of the number of derived mutations in the SFS are derived for a general Λ -coalescent.

Substantial theoretical research has been undertaken to identify the asymptotic behavior of the frequency spectrum associated with Λ -coalescents, culminating in the recent convergence results, as sample size tends to infinity, contained in Berestycki *et al.* (2013). However, so far it is not known how large samples need to be for the spectrum to be well approximated by the limiting asymptotic results of Berestycki *et al.* (2013) or, indeed, whether the approximation is equally good for all values of the relevant coalescent parameter. This question is addressed using simulations.

Finally, as an illustration we apply our methods to site-frequency spectra obtained for Atlantic cod (Árnason *et al.* 2000; Sigurgíslason and Árnason 2003; Árnason 2004) and fit Λ -coalescents to the data. The problem of distinguishing between coalescent classes by means of observing a realization of the SFS is also addressed using simulations.

Setup and Previously Known Results

Notation

A formal definition of the SFS is our starting point. We assume the *infinitely many-sites model* (Kimura 1969; Watterson 1975) of population genetics, where the genealogy of a sample of size n is described by an n - Λ -coalescent. Mutations occur at a rate $\theta/2$ on the coalescent branches. Suppose further that we can distinguish between mutant and wild type, *e.g.*, with the help of an outgroup. Then, the SFS of an n sample is given by

$$\xi^{(n)} := (\xi_1^{(n)}, \dots, \xi_{n-1}^{(n)}),$$

where the $\xi_i^{(n)}$, $i = 1, \dots, n-1$ denote the number of sites at which variants appear i times in our sample. The $\xi_i^{(n)}$ are also referred to as the number of *derived* mutations in i copies, when the ancestral state is known. The vector $\xi^{(n)}$ is called the *unfolded* site-frequency spectrum of the n sample.

In the case where the ancestral state (variant) is unknown, one often considers $\eta_i^{(n)} = \xi_i^{(n)}$ for $i = 1, \dots, \lfloor n/2 \rfloor$, which is the number of sites at which the less frequent variant (the minor allele) is present in i sequences, and $\lfloor n/2 \rfloor := \mathbf{1}_{(n \text{ even})}n/2 + \mathbf{1}_{(n \text{ odd})}(n-1)/2$, where $\mathbf{1}_{(A)} = 1$ if event A is true and zero otherwise. The observed polymorphisms become “few and far between” with corresponding small expected values as i increases, at least for the Atlantic cod data we investigate. Thus, for $i > \hat{i}$ for some suitable $\hat{i}$, we lump them together into one class labeled as $\hat{i}_+$. Considering only the first few classes and lumping the rest also speeds up the computations.

Figure 1 Illustration of the occurrence of an event with probability $p^{(n)}[k, b]$, with $k = 2$, and $2 \leq b = n - 3$. Ancestral lineage labeled as $(2, 1)$, present when there are only two active ancestral lineages, is ancestral to b leaves (b sampled sequences). In this example, the first merger in the genealogical history of the n leaves is a 3 merger not involving any of the encircled b leaves. The second event is a b merger with b edges subtending the encircled b leaves merging to block $(2, 1)$ (square). A “level” refers to the values of the lineage-counting process Y_t .

Then, $\boldsymbol{\eta}^{(n)} := (\eta_1^{(n)}, \dots, \eta_{(n/2)}^{(n)})$ is called the *folded* site-frequency spectrum of an n sample.

A Λ -coalescent is a partition-valued (one can think of enumerated DNA sequences and their ancestral relationships) exchangeable coalescent process determined by a finite measure Λ on $[0, 1]$ (Donnelly and Kurtz 1999; Pitman 1999; Sagitov 1999). For a comprehensive overview see, e.g., Berestycki (2009). If there are currently n blocks in the partition (i.e., n active ancestral lineages), any particular set of k blocks merges into one at rate

$$\lambda_{n,k} = \int_{[0,1]} x^{k-2} (1-x)^{n-k} \Lambda(dx), \quad k = 2, \dots, n. \quad (1)$$

Certain special cases of a Λ -coalescent include the following, some of which we use in our examples below:

One recovers the Kingman coalescent for

$$\Lambda(dx) = \delta_0(dx), \quad (2)$$

i.e. when the measure Λ is concentrated at the point 0.

The Beta($2 - \alpha, \alpha$)-coalescent (Schweinsberg 2003) is a Λ -coalescent characterized by the Λ -measure

$$\Lambda(dx) = \frac{\Gamma(2)}{\Gamma(2-\alpha)\Gamma(\alpha)} x^{1-\alpha} (1-x)^{\alpha-1} dx, \quad (3)$$

with $\alpha \in (0, 2)$, i.e., when the measure Λ is associated with the Beta-distribution with parameters $2 - \alpha$ and α . The limiting case $\alpha = 2$ (in the sense of weak convergence of measures) corresponds to the Kingman coalescent.

Eldon and Wakeley (2006) considered purely atomic Λ -coalescents of the types

$$\Lambda(dx) = \delta_\psi(dx) \quad (4)$$

and

$$\Lambda(dx) = \frac{2}{2+\psi^2} \delta_0(dx) + \frac{\psi^2}{2+\psi^2} \delta_\psi(dx), \quad (5)$$

with $\psi \in [0, 1]$, where $\psi = 0$ gives the Kingman coalescent.

Figure 2 Illustration of the occurrence of the event with probability $p_{\text{eq}}^{(n)}[k; i, j]$ with $i = j = 1$. The ancestral lines $(k, 1)$ and $(k, 2)$ are each ancestral to one leaf. Level k is reached in one $n - k + 1$ merger not involving the two leaves subtended by $(k, 1)$ or $(k, 2)$.

In the application examples that follow, we are mainly concerned with coalescents (3) and (4). Thus, the parameter α ($1 < \alpha \leq 2$) refers exclusively to coalescent (3), and the parameter ψ ($0 < \psi < 1$) to coalescent (4).

For α , we focus our attention on the parameter range $1 < \alpha < 2$, since this corresponds to Beta-coalescents that “come down from infinity”, i.e., admit a unique most-recent common ancestor even in the infinite-sample limit (Pitman 1999; Schweinsberg 2000). It can be argued whether this property is necessary from a biological point of view. However, our empirical results for real data sets yield so far only estimates of α between 1 and 2. In addition, the coalescent process derived from Schweinsberg’s (2003) population model converges to coalescent (3) only when $1 \leq \alpha < 2$. For $0 < \alpha < 1$, the coalescent process is a discrete-time Ξ -coalescent (Schweinsberg 2003).

Related asymptotic results

In inference, one often relies on asymptotic results. The asymptotic behavior of the site- and allele-frequency spectrum of Λ -coalescents for (very) large sample sizes (as $n \rightarrow \infty$), as well as related asymptotic results concerning the distributions of tree lengths or the “speed of coming down from infinity”, i.e., when the sample size is infinite (a mathematical peculiarity), how long it takes the ancestral lineages to become finite in number, is an active field of mathematical research. The concept of coming down from infinity may yet have real implications for inference. By way of example, the point-mass coalescent (4) does not come down from infinity, which implies that external branches become more dominating as sample size increases, regardless of the value of ψ . On the other hand, the Beta-coalescent (3) does come down from infinity, as does the Kingman coalescent.

Important asymptotic results regarding the SFS are obtained by Berestycki *et al.*, (2013, Theorem 3), who provide convergence results (in the almost sure sense), as sample size tends to infinity, of the scaled SFS for an important class of Λ -coalescents, including the Beta-coalescent (3), as is now explained. We say that the measure Λ has (strong) α -regular variation at zero if

$$\Lambda(dx) = f(x)dx, \quad \text{where } f(x) \sim Ax^{1-\alpha} \text{ as } x \rightarrow 0,$$

for some $\alpha \in (1, 2)$ and $A > 0$. This holds for the Beta($2 - \alpha, \alpha$)-coalescent from (3). Then, the almost sure asymptotic

Figure 3 Illustration of the occurrence of the event with probability $p_{un}^{(n)}[k, i; \ell, j]$. Edge $(\ell, 1)$ in level ℓ subtends j leaves, while edge $(k, 1)$ in level k subtends i leaves; $2 \leq k < \ell \leq n$, and the two encircled groups of i and j leaves are distinct. Level ℓ is reached in one merger including all of the j leaves. The following merger includes all of the i leaves.

behavior of each $\xi_i^{(n)}$ as $n \rightarrow \infty$ is determined by the behavior of $f(x)$ as $x \downarrow 0$ for a large class of Λ -coalescents.

Theorem 1 (Berestycki *et al.* 2013). Assume that Λ has (strong) α -regular variation at zero for some $1 < \alpha < 2$ and $A > 0$. Then, for the site-frequency spectrum of an n sample, where the underlying genealogy is governed by a Λ -coalescent and mutation rate $\theta/2$,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\xi_i^{(n)}}{n^{2-\alpha}} &= \frac{\theta}{2} C_{A,\alpha} \frac{(2-\alpha)\Gamma(i+\alpha-2)}{i!\Gamma(\alpha-1)} \\ &= \frac{\theta}{2} C_{A,\alpha} (2-\alpha) \frac{(\alpha-1) \cdots (\alpha+i-3)}{i!}, \end{aligned}$$

in the almost sure sense, for $i = 1, \dots, n$, where

$$C_{A,\alpha} = \frac{\alpha(\alpha-1)}{A\Gamma(2-\alpha)(2-\alpha)}.$$

We formulate **Theorem 1** with a mutation rate $\theta/2$ instead of θ as in Berestycki *et al.* (2013) to be consistent with Fu (1995) and other literature. The asymptotic results should be of direct relevance for us in the case of large sample sizes. However, the rate of convergence of the asymptotic results in terms of sample size n and the parameter α seems to be unresolved. The issue of rate of convergence is investigated below, using simulations.

From a statistical point of view an ideal result would be an (asymptotic) result for the joint distribution of the normalized SFS for general Λ -coalescents, and this may well soon be within reach. For the Kingman case, this has already been achieved: There is asymptotic normality for the joint distribution of tree lengths. Kersting and Stanciu (2013) obtain an asymptotic normality result for the scaled branch lengths associated with the Kingman coalescent, from which Kersting and Stanciu deduce that the asymptotic joint distribution of the frequency spectrum is independent Poisson.

As a partial analog, Kersting *et al.* (2013) prove convergence of the total length of external branches of the Beta-coalescent (3) in distribution to a stable random variable with index α . Their result indicates that the joint distribution of the branch lengths associated with a Λ -coalescent will not be asymptotically normal, but might instead be a multivariate stable distribution. However, the rate of convergence is not known in the results of Kersting *et al.* (2013).

Figure 4 Illustration of the occurrence of the event with probability $p_{ne}^{(n)}[k, i; \ell, j]$. Edge $(\ell, 1)$ in level ℓ subtends the encircled subset of j leaves, while edge $(k, 1)$ in level k subtends $(\ell, 1)$ and the larger encircled set of i leaves necessarily containing the set of j leaves.

The Expected Value, Variance, and Covariance of Derived Mutations in the SFS Associated with a Λ -Coalescent

In this section we compute the expected value $\mathbb{E}[\xi_i^{(n)}]$, the variance $\text{Var}(\xi_i^{(n)})$, and the covariance $\text{Cov}(\xi_i^{(n)}, \xi_j^{(n)})$ of the SFS associated with Λ -coalescents. While exact solutions for a finite sample size n appear quite hard to obtain due to the multiple-merger property of Λ -coalescents, we derive recursions for these quantities. Fu (1995) (cf. Theorem S1 in Supporting Information) obtained closed-form expressions for these quantities when associated with the Kingman coalescent.

As above, let $\xi_i^{(n)}$ denote the random number of derived mutations in i copies in an n sample with mutation rate $\theta/2$ and genealogy governed by a Λ -coalescent. Before stating our first result, we recall some simple properties of the block-counting process associated with a Λ -coalescent from Birkner and Blath (2008). The block-counting process (Y_t) simply counts the number of ancestral lineages present each time. Recall the corresponding rates $\lambda_{n,k}$ from (1). When we refer to number of “leaves” (n), we mean the initial sample size. Thus, a “leaf” refers to one sampled (DNA) sequence. By using leaves rather than sequences we emphasize that we are thinking of the sequences as vertices in a graph describing the ancestral relations of the leaves. Let $(Y_t)_{t \geq 0}$ be the block-counting process of our Λ -coalescent, which is a continuous-time Markov chain on $\mathbb{N}$ with jump rates

$$q_{ij} = \binom{i}{i-j+1} \lambda_{i,i-j+1}, \quad i > j \geq 1.$$

The total jump rate away from state i is $-q_{ii} = \sum_{j=1}^{i-1} q_{ij}$. By

$$p_{ij} := \frac{q_{ij}}{-q_{ii}} \quad (6)$$

we denote the transition probabilities of the embedded discrete skeleton chain. Let

$$g(n, m) := \mathbb{E}_n \left[\int_0^\infty \mathbf{1}_{(Y_s=m)} ds \right] \quad \text{for } n \geq m \geq 2 \quad (7)$$

Figure 5 The covariance $\text{Cov}(B_i^{(n)}, B_j^{(n)})(C)$ for sample size $n = 40$ for the $\text{Beta}(2 - \alpha, \alpha)$ -coalescent with α as shown. The bottom panels show only the covariance terms, with the variance terms $\text{Var}(B_i^{(n)})$ set equal to zero. The covariances range between -0.094 and 14.85 when $\alpha = 1.05$ and between -0.045 and 2.06 when $\alpha = 1.5$. One has $\text{Cov}(\xi_i^{(n)}, \xi_j^{(n)}) = (\theta^2/4)\text{Cov}(B_i^{(n)}, B_j^{(n)})$ for $i \neq j$, and $\text{Var}(\xi_i^{(n)}) = (\theta/2)\mathbb{E}[B_i^{(n)}] + (\theta^2/4)\text{Var}(B_i^{(n)})$.

be the expected amount of time that Y_t , starting from n , spends in m . Decomposing according to the first jump of Y_t gives a recursion for $g(n, m)$,

$$g(n, m) = \sum_{k=m}^{n-1} p_{nk} g(k, m), \quad n > m \geq 2, \quad (8)$$

and

$$g(m, m) = \frac{1}{-q_{mm}}, \quad m \geq 2.$$

The expected value $\mathbb{E}[\xi_i^{(n)}]$ of derived mutations

For $1 < k \leq n$, as in Fu (1995), an important quantity is the probability

$$p^{(n)}[k, b], \quad \text{for } b \in \{1, \dots, n - (k - 1)\},$$

that in a Λ -coalescent starting from n leaves, conditioned that there are at some point in time exactly k branches, a given one of these k branches (e.g., the first, if we think of some ordering) subtends exactly b leaves (see Figure 1).

A recursion for $p^{(n)}[k, b]$ (Equation A2), given in the Appendix (Proposition A1), can be solved numerically, separately for each k . Equation (A2) allows us to derive a recursion for the expected frequency spectrum associated with Λ -coalescents, given in the following proposition.

Proposition 1. Under the above assumptions, we have, for $i = 1, \dots, n - 1$,

$$\mathbb{E}[\xi_i^{(n)}] = \frac{\theta}{2} \sum_{k=2}^{n-i+1} p^{(n)}[k, i] \cdot k \cdot g(n, k).$$

The proof proceeds by decomposing according to the “level” (the values taken by the lineage-counting process Y_t) of the coalescent tree where a mutation arises—see the Appendix for details. **Proposition 1** and its proof are a natural extension of the arguments of Griffiths and Tavaré (1998) to the multiple-merger case.

The expression in **Proposition 1** can be used to define a quantity $(\varphi_n(i))$, the “expected normalized frequency spectrum”, which describes the shape of the genealogy and does not depend on θ , as follows. Let $B_i^{(n)}$ denote the random total length of branches subtending i leaves, for $1 \leq i < n$. Then, clearly, $\mathbb{E}[\xi_i^{(n)}] = (\theta/2)\mathbb{E}[B_i^{(n)}]$, with

$$\mathbb{E}[B_i^{(n)}] = \sum_{k=2}^{n-i+1} p^{(n)}[k, i] \cdot k \cdot g(n, k).$$

If we now let $B^{(n)}$ denote the random total length of the genealogy started from n leaves, and by $\xi^{(n)} = \xi_1^{(n)} + \dots + \xi_{n-1}^{(n)}$ we denote the random total number of segregating sites in a sample of size n gene copies (DNA sequences), we have

Figure 6 The covariance $\text{Cov}(B_i^{(n)}, B_j^{(n)})$ (C) for sample size $n = 40$ for the Beta($2 - \alpha, \alpha$)-coalescent with α as shown and with covariances $\text{Cov}(B_i^{(n)}, B_{n-i}^{(n)})$ and the variance terms set equal to zero.

$$\mathbb{E}[\xi^{(n)}] = (\theta/2)\mathbb{E}[B^{(n)}].$$

If T_ℓ denotes the random time during which there are ℓ active lineages, for $2 \leq \ell \leq n$, we have

$$\mathbb{E}[B^{(n)}] = \sum_{\ell=2}^n \ell \mathbb{E}[T_\ell].$$

We define φ as

$$\varphi_n(i) := \frac{\mathbb{E}[\xi_i^{(n)}]}{\sum_{\ell=2}^n \ell \mathbb{E}[T_\ell]} = \frac{\sum_{k=2}^{n-i+1} p^{(n)}[k, i] \cdot k \cdot g(n, k)}{\sum_{\ell=2}^n \ell g(n, \ell)}. \quad (9)$$

Loosely speaking, the probability that a mutation, under the infinitely many-sites assumption (Kimura 1969; Watterson 1975), with known ancestral types, appears i times in a sample of size n is $\varphi_n(i)$. The quantity $\varphi_n(i)$ is a quotient of expectations that can be interpreted as size biasing—it should not be confused with averaging over the quotient of these quantities, e.g., obtained from an empirical study, which has a different interpretation and will, at least for finite sample sizes, take different values. However, one can apply $\varphi_n(i)$ as the main ingredient in a pseudolikelihood approach to infer coalescence parameters, as we do below.

The covariance $\text{Cov}(\xi_i^{(n)}, \xi_j^{(n)})$ of derived mutations

To compute covariances one needs to compute the probability that two edges (ancestral lineages) drawn at random without replacement (ordered sampling) from the past, say from Y -states k and $\ell \leq n$, produce i (resp.) j subtended leaves. To this end, we need to distinguish three situations, namely whether our two edges are taken from the same “layer” (that is, when Y is in the same state k for both edges) or, if this is not the case, whether one of our two edges is a descendant of the other (“nested case”) or not (“unnested case”). In our calculations the leaves are *unlabeled*.

Case 1: Two edges in the same layer: For $2 \leq k \leq n$ let $p_{\text{eq}}^{(n)}[k; i, j]$ be the probability that in an n - Λ -coalescent, conditioned on there being $k \geq 2$ ancestral lineages at some

time in the past, if we draw two of the k lines at random without replacement from these k (ordered sample), the first drawn line has i and the second j subtended leaves (see Figure 2). A recursion for $p_{\text{eq}}^{(n)}$ (Proposition A2) and proof are given in the *Appendix*.

Case 2: Two edges in different layers, unnested case: For $2 \leq k < \ell \leq n$ let $p_{\text{un}}^{(n)}[k, i; \ell, j]$ be the probability that in an n - Λ -coalescent, conditioned on there being k lines at some time in the past and ℓ lines at some (other) time in the past, if we draw one of the k lines at random and independently one of the ℓ lines at random (ordered sample), the first drawn line has i and the second j subtended leaves, and the second is not a descendant of the first (see Figure 3). A recursion for $p_{\text{un}}^{(n)}$ (Proposition A3) is obtained in a similar way to that for $p_{\text{eq}}^{(n)}$, and is given in the *Appendix* along with a proof.

Case 3: Two edges in different layers, nested case: For $2 \leq k < \ell \leq n$ let $p_{\text{ne}}^{(n)}[k, i; \ell, j]$ be the probability that in an n - Λ -coalescent, conditioned on there being k lines at some time in the past and $\ell > k$ lines at some more recent time in the past, if we draw (ordered sample) one of the k lines at random $[(k, 1)$, say, see Figure 4] and independently one of the ℓ lines $[(\ell, 1)$, say, see Figure 4] at random, the first drawn line has i and the second j subtended leaves, and the second $(\ell, 1)$ is a descendant of the first $[(k, 1)$, so necessarily $i \geq j$, and otherwise the probability is zero; see Figure 4]. The recursion for $p_{\text{ne}}^{(n)}$ is given in Proposition A4 in the *Appendix*, with the proof following similar arguments to proofs of Propositions A2 and A3.

Unfortunately, the computational complexity of the recursions for $p_{\text{eq}}^{(n)}$ (Equation A4), $p_{\text{un}}^{(n)}$ (Equation A6), and $p_{\text{ne}}^{(n)}$ (Equation A8), as they stand, is $O(n^5)$, thus restricting numerical computations to relatively small values of n .

Second moments and covariance of the frequency spectrum: Now we can put our recursions for $p_{\text{eq}}^{(n)}$, $p_{\text{un}}^{(n)}$, and $p_{\text{ne}}^{(n)}$ together to obtain the covariances $\text{Cov}(\xi_i^{(n)}, \xi_j^{(n)})$. Our main result on the site-frequency spectrum associated with

Figure 7 The results of Berestycki *et al.* (2013) (**Theorem 1**) regarding convergence of the scaled unfolded frequency spectrum (shaded bars) compared to simulations (open bars) varying over α as shown, with $\theta = 1$. The left column is for sample size $n = 100$, and the right column is for sample size $n = 250$. The whiskers represent standard deviation.

Λ -coalescents, an expression for $\mathbb{E}[\xi_i^{(n)} \xi_j^{(n)}]$ in terms of $p_{\text{eq}}^{(n)}$, $p_{\text{un}}^{(n)}$, $p_{\text{ne}}^{(n)}$, $p^{(n)}[k, i]$, and $g(n, m)$, is now stated—a proof is in the Appendix.

Theorem 2. For $1 \leq i, j < n$, $2 \leq i + j \leq n$, we have

$$\begin{aligned} & \mathbb{E}[\xi_i^{(n)} \xi_j^{(n)}] \\ &= \frac{\theta^2}{4} \sum_{k=2}^n k(k-1) p_{\text{eq}}^{(n)}[k, i, j] \frac{g(n, k)}{g(k, k)} \frac{2}{(-q_{kk})^2} \\ &+ \mathbf{1}_{(i=j)} \sum_{k=2}^n k p^{(n)}[k, i] \frac{g(n, k)}{g(k, k)} \left(\frac{\theta}{2} \frac{1}{-q_{kk}} + \frac{\theta^2}{4} \frac{2}{(-q_{kk})^2} \right) \\ &+ \frac{\theta^2}{4} \sum_{k=3}^n \sum_{\ell=2}^{k-1} k k' \\ &\cdot \frac{p_{\text{un}}^{(n)}[k, i; \ell, j] + p_{\text{ne}}^{(n)}[k, i; \ell, j] + p_{\text{un}}^{(n)}[k, j; \ell, i] + p_{\text{ne}}^{(n)}[k, j; \ell, i]}{(-q_{kk})(-q_{\ell\ell})} \\ &\cdot \frac{g(n, \ell)}{g(\ell, \ell)} \frac{g(\ell, k)}{g(k, k)}. \end{aligned}$$

In Figure 5 and Figure 6 we graph the covariances (Equation A17) of the branch lengths $B_i^{(n)}$ and $B_j^{(n)}$, where $B_i^{(n)}$ denotes the total length of branches subtending i of n leaves, and the coalescent process is the Beta($2-\alpha$)-coalescent. The variances and covariances become completely dominated by the variance of the length of external branches when multiple mergers are more prominent in the genealogy, *i.e.*, when α is close to 1 (Figure 5). This means that the variances and covariances of the SFS will be dominated by the variance in the number of singletons, when associated with Λ -coalescents, and multiple mergers dominate the

genealogy, as would occur in a high-fecundity population with frequent occurrence of large offspring number events. In the Beta-coalescent (Equation 3), large offspring number events become more prominent as α approaches 1 and less prominent as α approaches 2. And even when α is more “modest” at 1.5, the covariances are still quite small compared to the variances. Figure 6 shows the covariances on their own scale, *i.e.*, with the variances and the diagonal covariances $\text{Cov}(B_i^{(n)}, B_{n-i}^{(n)})$ set at zero. These results suggest that one may model the site-frequency spectrum of an n - Λ -coalescent as composed of uncorrelated (but not independent) components $\xi_i^{(n)}$ for n large enough.

For fixed n , the covariances between $B_i^{(n)}$ and $B_{n-i}^{(n)}$ seem to be larger than the covariances off the diagonals. As for the Kingman case, this can be explained by considering the first branching event in the coalescent history forward in time after the root. This has substantial probability to be a binary split, say in subfamilies of size i and $n-i$, which yields a positive correlation between $B_i^{(n)}$ and $B_{n-i}^{(n)}$.

Extending our results to the case of Ξ -coalescents and multiple loci is a natural follow-up task, but lies beyond the scope of the current discussion and will be part of future research.

Simulation Studies and Data Analysis

Reliability of the asymptotic results of Berestycki *et al.* (2013)

The asymptotic results obtained by Berestycki *et al.* (2013) (see **Theorem 1**) can be used to obtain estimates of α provided the asymptotic results are “close” to the true value and

Figure 8 The asymptotic results of Berestycki *et al.* (2013) (**Theorem 1**) for the scaled singletons (lines) compared to the corresponding scaled exact expected values (symbols) as a function of sample size $n \in \{10, 20, 50, 100, 200, 300, 500, 1000, 2000, 10,000\}$ on log-scale and varying over α as shown. Thus, the solid line and circles refer to $\alpha = 1.05$, the short-dashed line and squares refer to $\alpha = 1.25$, and so on.

one has an estimate of θ . Ideally, such inference should take into account correlation between loci. Now we address the question of how large the sample must be for the SFS to be well approximated by the asymptotic results of Berestycki *et al.* (2013) and, indeed, whether one can trust the asymptotic results for all values of α .

Figure 7 compares the asymptotic results in **Theorem 1** to simulated values for sample size $n = 500$ (left column) and $n = 1000$ (right column) and with α varying as shown. Even though we have recursions to obtain the exact expected values for finite n , we were not able, with present computer power, to compute covariances for $n \geq 100$. The estimated expected values and the ones from the recursions are in agreement (results not shown). Even for $n = 10^4$ (Figure 1 and Figure 2), which would be a rather large sample, the asymptotic results (Figure 7, shaded bars) match the simulated ones (Figure 7, open bars) rather poorly when $\alpha < 1.5$, but the correspondence improves as α increases, and at $\alpha = 1.5$ the match is already quite good. For more common sample sizes $n \leq 10^3$ (Figure 3 and Figure 7) the match is not very good for $\alpha \leq 1.5$. Relying on asymptotic results to estimate α might thus give misleading results, even when sample size is quite large, as demonstrated in Figure 8, which compares the asymptotic results for the singletons to the scaled (with $n^{\alpha-2}$) exact expected values as sample size n and α vary as shown. The poor fit of the asymptotic results to exact values of the SFS, at least for lower values of α , suggests that exact results for the SFS, if only in the form of recursions, may be a better choice for inference methods.

Inference Methods and Data Analysis

Before we discuss our results on the analysis of the Atlantic cod data, we introduce our two inference methods, comparing observed and expected values using an ℓ^2 distance and a pseudolikelihood approach.

Table 1 Pseudo-log-likelihood values and maximum-likelihood estimates ($\hat{\alpha}$ and $\hat{\psi}$) of coalescent parameters α and ψ for the pseudolikelihood (11) with $p_i := \mathbb{E}[B_i^{(n)}]/\mathbb{E}[B^{(n)}]$

n	π	$\hat{\alpha}$	$\hat{\psi}$	$\log L(\hat{\alpha})$	$\log L(\hat{\psi})$	$\log L(0)$
10	0	1.96	0.02	-5.404	-5.405	-5.404
100		1.96	0.02	-19.027	-19.176	-19.021
1000		1.96	0.02	-38.868	-43.281	-38.847
10	1.05	1.06	0.30	-5.122	-5.159	-5.583
100		1.06	0.12	-26.179	-27.483	-35.3964
1000		1.04	0.04	-106.80	-120.525	-204.1335
10	1.5	1.50	0.14	-5.410	-5.419	-5.508
100		1.50	0.04	-23.731	-24.163	-25.272
1000		1.50	0.02	-71.143	-78.175	-81.3238
10	0.01	1.96	0.02	-54,110.92	-54,113.630	-54,113.59
100		1.90	0.02	-194,644.8	-194,884.834	-194,951.4
1000		1.62	0.02	-440,572.40	-432,453.967	-466,074.3
10	0.5	1.02	0.5	-19.219	-18.378	-23.28243
100		1.02	0.5	-85.749	-63.553	-207.8592
1000		1.02	0.5	-329.189	-130.413	-2,099.191

By $L(0)$ we denote the log-likelihood associated with the Kingman coalescent. Data are the expected unfolded frequency spectrum associated with a particular coalescent with parameter π . Values associated with the Λ_ψ -coalescent (Equation 4) were computed for the range $[0.02, 0.96]$, and those for Λ_α (Equation 3) in the range $[1.02, 1.96]$. One computes $\log L(0) = \sum_{i=1}^{n-1} (1/i) \log((2/i)/\mathbb{E}[L_K])$, $\mathbb{E}[L_K] = 2 \sum_{i=1}^{n-1} 1/i$.

The ℓ^2 distance

The simplest way to find a value of α or ψ that gives the best fit to the data would be to compare the expected values of the frequency spectrum to the observed one and report the parameter value that minimizes the sum of squares (Equation 10). To avoid having to jointly estimate the mutation rate, we can define the scaled frequency spectrum $\zeta_i^{(n)}$ as

$$\zeta_i^{(n)} := \frac{\eta_i^{(n)}}{\sum_{i=1}^{\lfloor n/2 \rfloor} \eta_i^{(n)}}.$$

The observed $\zeta_i^{(n)}$ are compared to the ratio of expected values

$$r_i^{(n)} := \frac{\mathbb{E}[\eta_i^{(n)}]}{\sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{E}[\eta_i^{(n)}]}.$$

Although clearly one should compare $\zeta_i^{(n)}$ to its actual expected value, simulation results indicate that our approximation is quite good (results not shown). A natural distance measure to compare $\zeta_i^{(n)}$ and $r_i^{(n)}$ is the ℓ^2 distance

$$\ell^2 = \sqrt{\sum_{i=1}^{n/2} (\zeta_i - r_i)^2}. \quad (10)$$

A similar distance measure is the G_ξ statistic proposed by Fu (1996), in which each term in the sum is weighted by the variance of $\xi_i^{(n)}$ ($\eta_i^{(n)}$). The computational cost of computing the variances for a Λ -coalescent prevents us from considering an identical statistic for Λ -coalescents. A drawback of Fu's (1996) G_ξ statistic is that it depends on the mutation

Table 2 The folded frequency spectrum for the Atlantic cod data in Árnason *et al.* (2000), Sigurgíslason and Árnason (2003), and Árnason (2004)

Source	s	η_1	η_2	η_3	η_4	η_5	η_6	η_7	η_8	η_9	η_{10}	η_{11+}
Árnason <i>et al.</i> (2000)												
Greenland (78)	9	3	1	0	1	0	0	1	1	1	0	1
Iceland (519)	23	12	1	4	1	0	0	0	0	0	0	5
Joint (597)	24	11	3	3	1	1	0	0	0	0	0	5
Sigurgíslason and Árnason (2003)												
Faroe Islands (74)	44	23	9	1	1	1	0	9	0	0	0	0
Árnason (2004) (1278)	39	17	4	4	2	4	1	0	1	0	0	6

The sample size of each data set is given in parentheses. Mutations of frequency classes from 11 onward are lumped together in the 11+ class except for the data from Faroe Islands, in which mutations from 7 onward are lumped into the 7+ class. The symbol s denotes the total number of segregating sites.

rate θ . Ideally, one would want to apply a statistic that did not depend on θ . Instead, a statistic that reflects the underlying genealogy is desirable, since the processes that one usually tries to learn about in population genetics affect the topology and the branch lengths of the genealogy of a sample in one way or another. Determining statistical significance levels of the ℓ^2 distance is postponed for now; however, one may apply the approach of Fu (1996) to determine significance levels and statistical power.

The pseudolikelihood approach

A more fundamental statistical inference procedure than simply applying the ℓ^2 statistic is a likelihood inference. Ideally, in our case, to construct a likelihood function, one would have a way of writing down the probability of observing a mutation i times in a sample in terms of the coalescence parameters. However, these probabilities depend on the coalescence parameters in a way that is hard to make explicit, and therefore we approximate them with the quantity $\varphi_n(i)$ (Equation 9). We interpret the observed site-frequency spectrum as an observation from a multinomial distribution, in which the probability of each class is estimated using $\varphi_n(i)$. For $1 \leq i < n$, $\varphi_n(i)$ is the expected number of mutations in i copies scaled by the expected total number of segregating sites, which corresponds to the expected total number of branches subtending i leaves scaled by the expected total size of the genealogy. Since we can compute $\varphi_n(i)$ only numerically, Fisher information cannot be computed easily, and it would not be *a priori* clear whether the variance and confidence bounds derived therefrom via classical statistical theory would be valid. Finally, we refer to our likelihood approach as a pseudolikelihood inference.

Now we state our pseudolikelihood function. Let s denote the total number of segregating sites, and let $\mathbf{s} = (s_1, \dots, s_k, s_{k+})$ be the observed folded frequency spectrum, where s_{k+} denotes the collapsed class. Write $\varphi_i := \varphi_n(i)$, and define $\varphi_{k+} = 1 - \varphi_1 - \dots - \varphi_k$. The pseudolikelihood function $L_s(\pi)$ takes a multinomial form, in which π denotes the appropriate coalescence parameter (ψ or α):

Figure 9 The ℓ^2 distance between ζ_i and r_i for the five different Atlantic cod data sets, with expected values obtained for the Beta-coalescent. The values associated with the Kingman coalescent are obtained for $\alpha = 2$. The A04 data set refers to the one in Árnason (2004).

$$L_s(\pi) = \binom{s}{s_1 \dots s_k s_{k+}} \varphi_1^{s_1} \dots \varphi_k^{s_k} \varphi_{k+}^{s_{k+}}. \quad (11)$$

The form of L can be seen from the way we view the data $\mathbf{s}$: as an observation from a multinomial distribution in which the probability of each class is estimated with φ_i . It suffices to consider the pseudo-log-likelihood function

$$l_s(\pi) = s_1 \log(\varphi_1) + \dots + s_{k+} \log(\varphi_{k+}).$$

As in the case of the statistics $\zeta_i^{(n)}$ and $r_i^{(n)}$, this pseudolikelihood approach does not require an estimate of θ . In Table 1 we do show that the pseudolikelihood performs as it should, when applied to the expected site-frequency spectrum of different coalescent processes.

Table 1 shows maximum-likelihood estimates of the coalescence parameters α and ψ , when the data are the expected values of the site-frequency spectrum, obtained for given values of the appropriate coalescent parameter, denoted by π in Table 1 and computed using our recursion (**Proposition 1**). Thus, $\pi = \alpha$ when $\pi \in (1, 2)$ and $\pi = \psi$ when $\pi \in (0, 1)$. For each combination of sample size n and π , the log-likelihood values are compared with the ones obtained for the Kingman coalescent $\log L(0)$. Highest log-likelihood values are always obtained for the correct parameter value even for sample size n as low as 10. And for modest sample sizes, one should be able to correctly identify the Kingman coalescent ($\pi = 0$ in Table 1).

In Figures 14–16 in supporting information, [File S1](#), we estimate the distribution of the maximum-likelihood estimator for different values of n , θ , and the coalescence parameters α and ψ . As one expects, the performance of the pseudolikelihood is better for larger values of n and θ .

A natural question to address is, How well do the probabilities $\varphi_n(i)$ (Equation 9) actually fit with the correct values $\mathbb{E}[\zeta_i^{(n)}]$? In Figures 5 and 6 in [File S1](#) we address this issue via simulations. The agreement between p_i and $\bar{R}_i$ is

Figure 10 The ℓ^2 distance between ζ_i and r_i for the five different Atlantic cod data sets, with expected values obtained for the point-mass coalescent. See Figure 9 for explanation of symbols.

best when the number n of leaves (sample size) is large and α is closer to 2 than to 1.

The C code written to carry out the computations is available upon request.

Atlantic cod mtDNA data analysis

An application of the recursion (**Proposition 1**) for the exact expected values for the site-frequency spectrum is now illustrated using data on Atlantic cod obtained from various localities in the North Atlantic (Árnason *et al.* 2000; Sigurgíslason and Árnason 2003; Árnason 2004). The shallow gene genealogies observed among the haplotypes sampled, coupled with very high fecundity of Atlantic cod, led Árnason (2004) to argue that the Kingman coalescent may not be an appropriate null model for Atlantic cod. Later studies (Eldon 2011; Steinrücken *et al.* 2013) support that conclusion by finding a better match between the data and the Beta-coalescent than obtained for the Kingman coalescent. Eldon (2011) employ the asymptotic results of Berestycki *et al.* (2013) on the site-frequency spectrum obtained from Árnason (2004), while Steinrücken *et al.* (2013) employ full-likelihood methods on the sequence data of Atlantic cod data sets of smaller size than in Árnason (2004).

The Atlantic cod mtDNA data sets we analyze are summarized in Table 2. The sequence data from the Faroe Islands (Sigurgíslason and Árnason 2003) consist of a 250-bp segment from the mitochondrial *cytochrome b* gene in addition to 242- and 74-bp segments for a total of 566 bp. The data from Greenland and Iceland described in Árnason *et al.* (2000) consist of the 250-bp segment and so do the combined data described in Árnason (2004).

In Figure 9 and Figure 10 the ℓ^2 distances comparing $\zeta_i^{(n)}$ and $r_i^{(n)}$ obtained for the different Atlantic cod data sets are compared. The corresponding parameter estimates are reported in Table 3. For the larger data sets one observes much “sharper” curves than for the smaller ones as one would expect, since the larger data sets carry more information. The difference in ℓ^2 distance between the coalescent processes is

Table 3 The parameter values minimizing the ℓ^2 distance

Data	$\hat{\alpha}$	$\hat{\psi}$	$\ell^2(\hat{\alpha})$	$\ell^2(\hat{\psi})$	$\ell^2(0)$
Faroe Islands	1.35	0.084	0.113	0.165	0.371
Greenland (G)	1.65	0.036	0.200	0.217	0.272
Iceland (I)	1.45	0.022	0.154	0.172	0.473
G + I	1.50	0.016	0.077	0.180	0.429
A04	1.50	0.012	0.121	0.278	0.569

The ℓ^2 distance $\ell^2(0)$ associated with the Kingman coalescent is given for reference. The reference “A04” refers to the large data set in Árnason (2004).

also biggest for the larger data sets, with the Beta-coalescent showing the best fit for all the data sets, although the difference between the two Λ -coalescents is not always large. However, based on this simple parameter estimation method, one would clearly prefer one of the two Λ -coalescents over the Kingman coalescent.

The maximum-likelihood estimates (Table 4) agree quite well with the ℓ^2 -distance approach in Table 3. Steinrücken *et al.* (2013) apply full-likelihood methods on small Atlantic cod data sets, including the Faroe Islands data set (Sigurgíslason and Árnason 2003) and the Greenland subsample of Árnason *et al.* (2000), to obtain estimates of α . Their estimates of $\hat{\alpha} = 1.5$ for the Greenland subsample and $\hat{\alpha} = 1.3$ for the Faroe Islands sample agree with our estimates, in particular the Faroe Islands sample.

The Beta-coalescent yields the highest likelihood for all the data sets (Table 4). The fit to the data set reported in Árnason (2004) is, in particular, much better than for the Kingman coalescent, as measured by the ℓ^2 distance. The superior fit of the Beta-coalescent is further illustrated in Figure 11, in which the expected site-frequency spectrum of the Beta-coalescent and the Kingman coalescent is compared to the observed one of Árnason (2004). A similar graph (Figure 4) for the Faroe Islands data of Sigurgíslason and Árnason (2003) is given in File S1. Comparing Akaike’s (1973) information criterion (Table 5) between models tells the same story (except for the Greenland data).

Discussion

Inferring coalescence parameters of Λ -coalescents from large samples is our main focus. To this end we derive recursions for the expected values and covariances of the site-frequency spectrum associated with a Λ -coalescent. The multiple-merger property of Λ -coalescents makes obtaining closed-form solutions, as done by Fu (1995) for the Kingman coalescent, prohibitively hard. The value of the recursions in obtaining point estimates of coalescent parameters for two specific Λ -coalescents is then illustrated in two ways, utilizing mtDNA data on Atlantic cod.

Asymptotic results play a key role in inference. Knowing for what regions of the parameter space the asymptotic results approximate the true values “well enough”, in some sense, in other words knowing the rate of convergence, is also quite important for inference. We investigate the rate of convergence for recent asymptotic results, as sample size $n \rightarrow \infty$,

Table 4 Estimates ($\hat{\alpha}$, $\hat{\psi}$) based on the pseudolikelihood (11) for the Atlantic cod data in Árnason *et al.* (2000), Sigurgíslason and Árnason (2003), and Árnason (2004)

Data	$\hat{\alpha}$	$\log L(\hat{\alpha})$	$\hat{\psi}$	$\log L(\hat{\psi})$	$\log L(0)$	$\hat{\theta}$	ℓ_B^2	ℓ_K^2
Faroe Islands	1.28	-82.239	0.06	-89.634	-97.543	4.577	4.730	16.441
Greenland (G)	1.70	-21.445	0.03	-21.615	-21.963	1.409	2.394	2.470
Iceland (I)	1.38	-29.895	0.02	-32.452	-40.547	0.982	3.683	10.880
G + I	1.47	-55.557	0.01	-60.403	-68.523	1.179	5.204	13.247
Árnason (2004)	1.48	-71.099	0.005	-81.846	-90.594	1.373	4.625	19.448

The estimate $\hat{\theta}$ of θ is the one associated with the Beta($2 - \alpha$, α)-coalescent. The maximum pseudo-log-likelihood values $\log L(\cdot)$, along with the corresponding value $\log L(0)$ associated with the Kingman coalescent, are given. The ℓ^2 distances between the observed and expected site-frequency spectrum computed for the maximum pseudolikelihood estimates are given for the Beta-coalescent (ℓ_B^2) and the Kingman coalescent (ℓ_K^2). The estimate $\hat{\theta}$ of θ is the one associated with the Beta-coalescent based on the total number of segregating sites.

of Berestycki *et al.* (2013) regarding the scaled site-frequency spectrum associated with Λ -coalescents, since we do not have an expression for the error in the asymptotic results. Our results show that the rate of convergence, in terms of sample size, is quite slow for low values of the coalescence parameter α . Thus, exact inference methods appear preferable.

A simple distance measure is employed on the scaled frequency spectrum, removing the need to jointly estimate the mutation rate. The observed spectrum is compared to the expected values for different values of the coalescence parameters, and the value that minimizes the distance is reported. A straightforward pseudolikelihood approach is also applied, where the probabilities of mutations to be in different classes are estimated from the recursions for the expected values of the spectrum. These two different approaches give very similar estimates for the two Λ -coalescents we consider. In addition, they are shown to be capable of distinguishing Λ -coalescents from the Kingman coalescent, as well as distinguishing between the two Λ -coalescents when sample size is not too small. Both approaches also show promise in distinguishing between different Λ -coalescents, and they do not require estimates of the mutation rate. Our results open up new possibilities to construct test statistics to test for “neutrality” when the underlying population model admits large offspring numbers, *i.e.*, in highly fecund natural populations such as Atlantic cod (Árnason 2004). One could, by way of example, consider linear weighted combinations of the site-frequency spectrum to construct tests for neutrality in high-fecundity populations, in the spirit of Achaz (2009). The present discussion does not address the statistical power of different statistics one could construct using our recursions and their performance under various scenarios, such as population expansion. This will be the subject of future work. Our aim here is to illustrate the wide applicability of the recursions we obtain for the mean and variance of the site-frequency spectrum associated with Λ -coalescents.

Λ -Coalescents are essentially single-locus models, as they admit only one merger each time. Models for multiple loci

Figure 11 The folded frequency spectrum (open bars) of the data of Árnason (2004) along with predictions of the Kingman coalescent (bars with light shading) and the Beta($2 - \hat{\alpha}$, $\hat{\alpha}$)-coalescent (bars with dark shading). The vertical lines represent the standard deviation, obtained for the Beta($2 - \hat{\alpha}$, $\hat{\alpha}$)-coalescent from 10^5 iterations. The class labeled “11” represents the collated tail of the spectrum, from 11 to $1278/2$.

would naturally involve Ξ -coalescents admitting simultaneous multiple mergers. Intuitively, one might think that multilocus inference methods would be better able to distinguish between different coalescent processes and, indeed, the underlying population models, than single-locus approaches. A natural question to address is the sample size, as well as the number of loci, one would recommend to identify the appropriate coalescent process. Such issues are truly relevant, not least in part by advances in DNA sequencing technology. Our partial response is that our results indicate that one may need only moderate sample sizes to distinguish between Λ -coalescents. The question of the number of loci required is still very much open, not least because the development of ancestral recombination graphs admitting simultaneous multiple mergers is only in its early stages (Eldon and Wakeley 2008; Birkner *et al.* 2013).

A key distinguishing feature of the site-frequency spectrum drawn from a large offspring number population is the excess of singletons compared to the spectrum one expects in an ordinary Wright–Fisher population. When data contain sequencing errors, some authors have proposed analyzing the data without the singletons, leading to corresponding test statistics that exclude the singletons (Achaz 2008). At present we do not address the issue of sequencing errors, but we point out that for many natural high-fecundity populations, and with advances in DNA sequencing techniques, this may not be an issue. However, the issue of sequencing errors remains an important subject for future analysis.

The role of mtDNA as a suitable genetic marker for studying population history in general has been subject to some criticism (Ballard and Whitlock 2004; Bazin *et al.* 2006; Balloux 2010). A single genetic marker may not be sufficient to infer population history, due to stochasticity in the effects of a population’s demography on the genome (Ballard and Whitlock 2004). Natural selection may be continuously acting

Table 5 Akaike's information criterion (AIC) for the five cod data sets

Data	AIC(α)	AIC(ψ)	AIC(0)
Faroe Islands	166.478	181.268	195.086
Greenland (G)	44.89	45.23	43.926
Iceland (I)	61.79	66.904	81.094
G + I	113.114	122.806	137.046
Árnason (2004)	144.198	165.692	181.188

By AIC(0) we denote the AIC associated with the Kingman coalescent.

on (at least) parts of the mitochondrial genome of Atlantic cod, thus biasing any inference made on supposedly “neutral” segments of the nonrecombining mitochondrial genome (Maynard Smith and Haigh 1974; Gillespie 2000; Durrett and Schweinsberg 2005). In addition, we have not compared the data to models of population growth, which may mimic the effects of high fecundity. However, overfishing of Atlantic cod may have significantly diminished cod stocks throughout its range in the North Atlantic, which suggests that population expansion may not be an appropriate model for Atlantic cod. In addition, as mtDNA is inherited maternally, and female cods may lay millions of eggs annually (May 1967; Árnason 2004), our fit with Λ -coalescents might represent sweepstakes reproduction. Whatever the main reasons for the observed patterns of genetic diversity observed (see Árnason 2004, in particular the discussion on pp. 1882 and 1883), we stress that our purpose is simply to illustrate our inference methods, rather than reach a firm conclusion on the main mechanism shaping mtDNA diversity in Atlantic cod.

The two specific Λ -coalescents we consider will certainly not (one hopes) be the only Λ -coalescents ever applied to highly fecund natural populations. Recent work (Huillet and Möhle 2011a,b; Möhle 2011) introduces specific Λ -coalescent processes not considered in the present discussion. It remains an important exercise to compare all those new processes to actual data. In conclusion, we emphasize that even though one finds that one particular coalescent process fits the data better than another one, this does not automatically mean that one has found the correct model. Biological knowledge and insight remain key factors in identifying appropriate population models for natural populations. By way of example, the haploid population models underlying the two Λ -coalescents we focus on—(4) and (3)—are very different. In the model underlying (3), the key assumption is about the distribution of the number of viable (potential) offspring each individual contributes to the population (Schweinsberg 2003). In the model underlying (4), a single parent contributes a fixed number of offspring in each reproduction event (Eldon and Wakeley 2006). Such assumptions about the population model must form an integral part of any inference study.

Acknowledgments

We thank two referees for insightful comments that improved the presentation. M. Birkner was supported in part by Deutsche Forschungsgemeinschaft (DFG) grant BI 1058/2-1.

J. Blath and B. Eldon were supported in part by DFG grant BL 1105/3-1.

Literature Cited

- Achaz, G., 2008 Testing for neutrality in samples with sequencing errors. *Genetics* 179: 1409–1424.
- Achaz, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183: 249–258.
- Akaike, H., 1973 Information theory and an extension of the maximum likelihood principle, pp. 267–281 in *Second International Symposium on Information Theory*, edited by B. N. Petrov and F. Czaki. Akademiai Kiadó, Budapest.
- Avise, J. C., R. M. Ball, and J. Arnold, 1988 Current vs. historical population sizes in vertebrate species with high gene flow: a comparison based on mitochondrial DNA lineages and inbreeding theory for neutral mutations. *Mol. Biol. Evol.* 5: 331–344.
- Árnason, E., 2004 Mitochondrial cytochrome *b* variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. *Genetics* 166: 1871–1885.
- Árnason, E., P. H. Petersen, K. Kristinsson, H. Sigurgíslason, and S. Pálsson, 2000 Mitochondrial cytochrome *b* DNA sequence variation of Atlantic cod from Iceland and Greenland. *J. Fish Biol.* 56: 409–430.
- Ballard, J. W. O., and M. C. Whitlock, 2004 The incomplete natural history of mitochondria. *Mol. Ecol.* 13: 729–744.
- Balloux, F., 2010 The worm in the fruit of the mitochondrial DNA tree. *Heredity* 104: 419–420.
- Barton, N. H., A. M. Etheridge, and A. Véber, 2010 A new model for evolution in a spatial continuum. *Electron. J. Probab.* 7: 162–216.
- Barton, N. H., A. M. Etheridge, and A. Véber, 2013 Modelling evolution in a spatial continuum. *J. Stat. Mech.* 2013: 1002.
- Bazin, E., S. Glemin, and N. Galtier, 2006 Population size does not influence mitochondrial genetic diversity in animals. *Science* 312: 570–572.
- Beckenbach, A. T., 1994 Mitochondrial haplotype frequencies in oysters: neutral alternatives to selection models, pp. 188–198 in *Non-Neutral Evolution*, edited by B. Golding. Chapman & Hall, New York.
- Berestycki, J., N. Berestycki, and V. Limic, 2013 Asymptotic sampling formulae for lambda-coalescents. *ArXiv*:1201.6512.
- Berestycki, N., 2009 Recent progress in coalescent theory. *Ensaio Mathématiques* 16: 1–193.
- Birkner, M., and J. Blath, 2008 Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *J. Math. Biol.* 57: 435–465.
- Birkner, M., and J. Blath, 2009 Measure-valued diffusions, general coalescents and population genetic inference, pp. 329–363 in *Trends in Stochastic Analysis*, edited by J. Blath, P. Mörters, and M. Scheutzow. Cambridge University Press, Cambridge.
- Birkner, M., J. Blath, M. Möhle, M. Steinrücken, and J. Tams, 2009 A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks. *ALEA Lat. Am. J. Probab. Math. Stat.* 6: 25–61.
- Birkner, M., J. Blath, and M. Steinrücken, 2011 Importance sampling for Lambda-coalescents in the infinitely many sites model. *Theor. Popul. Biol.* 79: 155–173.
- Birkner, M., J. Blath, and B. Eldon, 2013 An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics* 193: 255–290.
- Boudry, P., B. Collet, F. Cornette, V. Hervouet, and F. Bonhomme, 2002 High variance in reproductive success of the Pacific oyster (*Crassostrea Gigas*, Thunberg) revealed by microsatellite-based

- parentage analysis of multifactorial crosses. *Aquaculture* 204: 283–296.
- Coop, G., and P. Ralph, 2012 Patterns of neutral diversity under general models of selective sweeps. *Genetics* 192: 205–224.
- Donnelly, P., and T. G. Kurtz, 1999 Particle representations for measure-valued population models. *Ann. Probab.* 27: 166–205.
- Durrett, R., and J. Schweinsberg, 2004 Approximating selective sweeps. *Theor. Popul. Biol.* 66: 129–138.
- Durrett, R., and J. Schweinsberg, 2005 A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stoch. Proc. Appl.* 115: 1628–1657.
- Eldon, B., 2011 Estimation of parameters in large offspring number models and ratios of coalescence times. *Theor. Popul. Biol.* 80: 16–28.
- Eldon, B., and J. Wakeley, 2006 Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* 172: 2621–2633.
- Eldon, B., and J. Wakeley, 2008 Linkage disequilibrium under skewed offspring distribution among individuals in a population. *Genetics* 178: 1517–1532.
- Fu, Y., 1995 Statistical properties of segregating sites. *Theor. Popul. Biol.* 48: 172–197.
- Fu, Y., 1996 New statistical tests of neutrality for DNA samples from a population. *Genetics* 143: 557–570.
- Gillespie, J. H., 2000 Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* 155: 909–919.
- Griffiths, R. C., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. *Commun. Stat. Stoch. Models* 14: 273–295.
- Hedgecock, D., and A. I. Pudovkin, 2011 Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. *Bull. Mar. Sci.* 87: 971–1002.
- Hedrick, P., 2005 Large variance in reproductive success and the N_e/N ratio. *Evolution* 59: 1596–1599.
- Huillet, T., and M. Möhle, 2011a On the extended Moran model and its relation to coalescents with multiple collisions. *Theor. Popul. Biol.* 87: 5–14.
- Huillet, T., and M. Möhle, 2011b Population genetics models with skewed fertilities: forward and backward analysis. *Stoch. Models* 27: 521–554.
- Kersting, G., and I. Stanciu, 2013 The internal branch lengths of the Kingman coalescent. *ArXiv:1303.4562*.
- Kersting, G., I. Stanciu, and A. Wakolbinger, 2013 The total external branch length of beta-coalescents. *ArXiv:1212.6070*.
- Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61: 893–903.
- Kingman, J. F. C., 1982 The coalescent. *Stoch. Proc. Appl.* 13: 235–248.
- Li, G., and D. Hedgecock, 1998 Genetic heterogeneity, detected by PCR-SSCP, among samples of larval Pacific oysters (*Crassostrea gigas*) supports the hypothesis of large variance in reproductive success. *Can. J. Fish. Aquat. Sci.* 55: 1025–1033.
- Hedgecock, D., V. Chow, and R. S. Waples, 1992 Effective population numbers of shellfish broodstocks estimated from temporal variance in allelic frequencies. *Aquaculture* 108: 215–232.
- May, A. W., 1967 Fecundity of Atlantic cod. *J. Fish. Res. Board Can.* 24: 1531–1551.
- Maynard Smith, J., and J. Haigh, 1974 The hitchhiking effect of a favorable gene. *Genet. Res.* 23: 23–35.
- Möhle, M., 2011 Coalescent processes derived from some compound Poisson population models. *Electron. Comm. Probab.* 16: 567–582.
- Möhle, M., and S. Sagitov, 2001 A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.* 29: 1547–1562.
- Möhle, M., and S. Sagitov, 2003 Coalescent patterns in diploid exchangeable population models. *J. Math. Biol.* 47: 337–352.
- Myers, S., C. Fefferman, and N. Patterson, 2008 Can one learn history from the allelic spectrum? *Theor. Popul. Biol.* 73: 342–348.
- Pitman, J., 1999 Coalescents with multiple collisions. *Ann. Probab.* 27: 1870–1902.
- Planes, S., and P. Lenfant, 2002 Temporal change in the genetic structure between and within cohorts of a marine fish, *Diplodus sargus*, induced by a large variance in individual reproductive success. *Mol. Ecol.* 11: 1515–1524.
- Sagitov, S., 1999 The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* 36: 1116–1125.
- Sagitov, S., 2003 Convergence to the coalescent with simultaneous mergers. *J. Appl. Probab.* 40: 839–854.
- Sainudiin, R., K. Thornton, J. Hallow, J. Booth, M. Stillman *et al.*, 2011 Experiments with the site frequency spectrum. *Bull. Math. Biol.* 73: 829–872.
- Sargsyan, O., and J. Wakeley, 2008 A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor. Popul. Biol.* 74: 104–114.
- Schweinsberg, J., 2000 Coalescents with simultaneous multiple collisions. *Electron. J. Probab.* 5: 1–50.
- Schweinsberg, J., 2003 Coalescent processes obtained from supercritical Galton-Watson processes. *Stoch. Proc. Appl.* 106: 107–139.
- Sigurgíslason, H., and E. Árnason, 2003 Extent of mitochondrial DNA sequence variation in Atlantic cod from the Faroe Islands: a resolution of gene genealogy. *Heredity* 91: 557–564.
- Steinrücken, M., 2009 Multiple merger coalescents and population genetic inference. Ph.D. Thesis, Technische Universität, Berlin.
- Steinrücken, M., M. Birkner, and J. Blath, 2013 Analysis of DNA sequence variation within marine species using Beta-coalescents. *Theor. Popul. Biol.* 87: 15–24.
- Wakeley, J., 2007 *Coalescent Theory*. Roberts & Co, Greenwood Village, Colorado.
- Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7: 256–276.
- Williams, G. C., 1975 *Sex and Evolution*. Princeton University Press, Princeton, NJ.
- Živković, D., and T. Wiehe, 2008 Second order moments of segregating sites under variable population size. *Genetics* 180: 341–357.

Communicating editor: J. Hermisson

Appendix

A Recursion for $p^{(n)}[k, b]$

Assume we start the block-counting process in some state $Y_0 = n$ and denote its distribution by $\mathbb{P}_n$. Then, the probability that the first jump takes it to state $n' < n$, conditionally on the event that it hits state $k \leq n'$, is given by

$$\mathbb{P}_n\{Y \text{ first jumps from } n \text{ to } n' | Y \text{ hits } k\} = \frac{\mathbb{P}_n\{Y \text{ first jumps to } n' \cap Y \text{ hits } k\}}{\mathbb{P}_n\{Y \text{ hits } k\}} = p_{n,n'} \frac{\mathbb{P}_{n'}\{Y \text{ hits } k\}}{\mathbb{P}_n\{Y \text{ hits } k\}} = p_{n,n'} \frac{g(n', k)}{g(n, k)}, \quad (\text{A1})$$

due to the Markov property of Y . This well-known observation will be useful later.

Decomposing according to the first jump of Y , starting from n , yields the following recursion for $p^{(n)}[k, b]$.

Proposition A1. *For $1 < k \leq n$, we have*

$$p^{(n)}[k, b] = \sum_{n'=k}^{n-1} p_{n,n'} \frac{g(n', k)}{g(n, k)} \left(\mathbf{1}_{(b > n-n')} \frac{b - (n - n')}{n'} p^{(n')}[k, b - (n - n')] + \mathbf{1}_{(b < n')} \frac{n' - b}{n'} p^{(n')}[k, b] \right), \quad (\text{A2})$$

with the obvious boundary conditions $p^{(n)}[n, b] = \delta_{1b}$ and $p^{(n)}[k, b] = 0$ if $b > n - (k - 1)$.

The boundary conditions allow us to replace the second indicator function on the right-hand side (rhs) by $\mathbf{1}_{(b \leq n - (k - 1))}$. In the case of the Kingman coalescent [indicated by an additional superscript, “(K)”], $p^{(n)}[k, b]$ becomes

$$p^{(n)(K)}[k, b] = \mathbf{1}_{(b > 1)} \frac{b - 1}{n - 1} p^{(n-1)(K)}[k, b - 1] + \mathbf{1}_{(b < n-1)} \frac{n - 1 - b}{n - 1} p^{(n-1)(K)}[k, b]$$

and Fu (1995) obtains a closed-form solution: $p^{(n)(K)}[k, b] = \binom{n-b-1}{k-2} / \binom{n-1}{k-1}$. A proof of Equation (A2) now follows.

Proof. Conditionally on Y hitting k , the first jump can take Y to any $n' \in \{k, \dots, n-1\}$. The probability of seeing a jump from n to n' , conditionally on hitting k , has probability $p_{n,n'}(g(n', k)/g(n, k))$, by (A1). Then, thinking “forward in time from n' lineages”, either the initial $(n - n' + 1)$ split occurred to one of the [then necessarily $b - (n - n')$] lineages subtended to the one we are interested in or it occurred to one of the (then necessarily $n' - b$) others.

Proof of Proposition 1

Proof. Let $T_k := \int_0^\infty \mathbf{1}_{(Y_s=k)} ds$ be the length of the time interval during which there are k lineages (which is possibly 0), and consider, in an n - Λ -coalescent,

$$\psi_n(b) := \text{expected total length of all branches with } b \text{ subtended leaves.}$$

One now obtains

$$\mathbb{E}[\xi_i^{(n)}] = \frac{\theta}{2} \psi_n(i) = \frac{\theta}{2} \sum_{k=2}^{n-i+1} p^{(n)}[k, i] k \mathbb{E}[T_k] = \frac{\theta}{2} \sum_{k=2}^{n-i+1} p^{(n)}[k, i] k g(n, k). \quad (\text{A3})$$

A Recursion for $p_{\text{eq}}^{(n)}[k; i, j]$

Proposition A2. *For $2 \leq k < n$ and $i, j \geq 1$; $i + j \leq n - k + 2$, we have*

$$\begin{aligned} p_{\text{eq}}^{(n)}[k; i, j] &= \sum_{m=k}^{n-1} p_{n,m} \frac{g(m, k)}{g(n, k)} \left[\frac{i - (n - m)}{m} p_{\text{eq}}^{(m)}[k; i - (n - m), j] \mathbf{1}_{(i > n - m)} \right. \\ &\quad \left. + \frac{j - (n - m)}{m} p_{\text{eq}}^{(m)}[k; i, j - (n - m)] \mathbf{1}_{(j > n - m)} + \frac{m - i - j}{m} p_{\text{eq}}^{(m)}[k; i, j] \mathbf{1}_{(i+j > m)} \right], \end{aligned} \quad (\text{A4})$$

and (as boundary condition)

$$p_{\text{eq}}^{(n)}[n; i, j] = \mathbf{1}_{(i=j=1)}, \quad n \in \mathbb{N}. \quad (\text{A5})$$

Proof. Decompose according to the first jump of Y , conditionally on hitting k . As in the proof of Proposition A1, the probability of jumping to some state $m \in \{k, \dots, n-1\}$ is $p_{n,m}(g(m,k)/g(n,k))$, by (12). Thinking “forward in time from m lineages”, the initial $(n-m+1)$ split occurred to one of the $i-(n-m)$ lineages subtended from the first sampled edge, or to one of the $j-(n-m)$ lineages subtended from the second sampled edge, or to one of the $m-i-b$ others.

One can derive a similar recursion for “labeled” leaves, where the order of the sample plays a role. For example, if we distinguish between individuals in our sample, we obtain for $p_{\text{eq}}^{(n)}(k; i, j)$, where the round brackets indicate that we refer to specific sets of i resp. j labeled leaves, the recursion

$$p_{\text{eq}}^{(n)}(k; i, j) = \sum_{n'=k}^{n-1} p_{n,n'} \frac{1}{\binom{n}{n-n'+1}} \left[\binom{i}{n-n'+1} p_{\text{eq}}^{(n')}(k; i-(n-n'), j) \right. \\ \left. + \binom{j}{n-n'+1} p_{\text{eq}}^{(n')}(k; i, j-(n-n')) + \binom{n-(i+j)}{n-n'+1} p_{\text{eq}}^{(n')}(k; i, j) \right]$$

for $n > 2$ [noting that $p_{\text{eq}}^{(n)}(k; i, j) = 0$ if $i < 1$ or $j < 1$ or $i+j \geq n-k+1$].

A Recursion for $p_{\text{un}}^{(n)}[k, i, \ell, j]$

Proposition A3. For $2 \leq k < \ell \leq n$ we have

$$p_{\text{un}}^{(n)}[k, i, \ell, j] = \sum_{m=\ell}^{n-1} p_{n,m} \frac{g(m,\ell)}{g(n,\ell)} \left[\frac{i-(n-m)}{m} p_{\text{un}}^{(m)}[k, i-(n-m); \ell, j] \mathbf{1}_{(i>n-m)} \right. \\ \left. + \frac{j-(n-m)}{m} p_{\text{un}}^{(m)}[k, i, \ell, j-(n-m)] \mathbf{1}_{(j>n-m)} + \frac{m-i-j}{m} p_{\text{un}}^{(m)}[k, i, \ell, j] \mathbf{1}_{(m>i+j)} \right] \quad (\text{A6})$$

with boundary conditions

$$p_{\text{un}}^{(n)}[k, i, n, j] = \mathbf{1}_{(j=1)} p^{(n)}[k, i] \frac{n-i}{n}, \quad 2 \leq k < n, \quad 1 \leq i < n. \quad (\text{A7})$$

Proof. As for the previous recursions, we decompose according to the first jump of Y , conditionally on hitting ℓ . Due to the Markov property we do not need to condition on hitting k as well. This second conditioning is “hidden” in the $p_{\text{un}}^{(m)}$ terms and in particular in the boundary terms (Equation A7). The rest of the argument is as usual.

A Recursion for $p_{\text{ne}}^{(n)}[k, i, \ell, j]$

Proposition A4. For $2 \leq k < \ell \leq n$ we have

$$p_{\text{ne}}^{(n)}[k, i, \ell, j] = \sum_{m=\ell}^{n-1} p_{n,m} \frac{g(m,\ell)}{g(n,\ell)} \left[\frac{i-j-(n-m)}{m} p_{\text{ne}}^{(m)}[k, i-(n-m); \ell, j] \mathbf{1}_{(i-j>n-m)} \right. \\ \left. + \frac{j-(n-m)}{m} p_{\text{ne}}^{(m)}[k, i-(n-m); \ell, j-(n-m)] \mathbf{1}_{(j>n-m)} + \frac{m-i}{m} p_{\text{ne}}^{(m)}[k, i, \ell, j] \mathbf{1}_{(m>i)} \right] \quad (\text{A8})$$

with boundary conditions

$$p_{\text{ne}}^{(n)}(k, i, n, j) = \mathbf{1}_{(j=1)} p^{(n)}[k, i] \frac{i}{n}, \quad 2 \leq k < n, \quad 1 \leq i < n. \quad (\text{A9})$$

The proof follows from arguments similar to the previous proofs.

The recursions for $p_{\text{eq}}^{(n)}$ (Equation A4), $p_{\text{un}}^{(n)}$ (Equation A6), and $p_{\text{ne}}^{(n)}$ (Equation A8) are (strictly) recursive in n and can be solved numerically in a straightforward way, starting from their respective boundary conditions (A5), (A7), and (A9). For numerical checking of implementations it may be useful to note that by definition,

$$\sum_{i,j} p_{\text{eq}}^{(n)}(k; i, j) = 1 \quad \text{for all } 2 \leq k \leq n, \quad (\text{A10})$$

$$\sum_{i,j} \left(p_{\text{un}}^{(n)}(k, i; \ell, j) + p_{\text{ne}}^{(n)}(k, i; \ell, j) \right) = 1 \quad \text{for all } 2 \leq k < \ell \leq n. \quad (\text{A11})$$

Proof of Theorem 2

Consider an n - Λ -coalescent, and let, for $2 \leq k \leq n$, $\ell \in [k]$,

$$L_{k,\ell}^{(n)} = \begin{cases} \text{No. leaves subtended to the } \ell\text{th edge while } k \text{ blocks (in } n\text{-coalescent),} \\ 0 \text{ if } k \text{ blocks are never realized.} \end{cases} \quad (\text{A12})$$

Now we think of labeled lines, as opposed to the unlabeled lines in the previous recursions. By symmetry, $\mathbb{P}\{L_{k,\ell}^{(n)} = i\}$ does not depend on $\ell \in [k] := \{1, \dots, k\}$. In fact

$$\mathbb{P}\{L_{k,\ell}^{(n)} = i\} = p^{(n)}[k, i] \frac{g(n, k)}{g(k, k)}, \quad \text{for } 2 \leq k \leq n, \ell \in [k]. \quad (\text{A13})$$

The factor $g(n, k)/g(k, k)$ is the probability that the block-counting process Y , starting from n lines, ever hits the state of k lines and thus removes the conditioning from the definition of $p^{(n)}[k, \cdot]$. Similarly, for $2 \leq k \leq n$, $\ell, \ell' \in [k]$, $\ell \neq \ell'$,

$$\mathbb{P}\{L_{k,\ell}^{(n)} = i, L_{k,\ell'}^{(n)} = j\} = \mathbb{P}\{L_{k,1}^{(n)} = i, L_{k,2}^{(n)} = j\} = p_{\text{eq}}^{(n)}[k; i, j] \frac{g(n, k)}{g(k, k)}, \quad (\text{A14})$$

and, for $2 \leq k < k' \leq n$, $\ell \in [k]$, $\ell' \in [k']$,

$$\mathbb{P}\{L_{k,\ell}^{(n)} = i, L_{k',\ell'}^{(n)} = j\} = \mathbb{P}\{L_{k,1}^{(n)} = i, L_{k',1}^{(n)} = j\} = \left(p_{\text{un}}^{(n)}(k, i; k', j) + p_{\text{ne}}^{(n)}[k, i; k', j] \right) \frac{g(n, k')}{g(k', k')} \frac{g(k', k)}{g(k, k)}. \quad (\text{A15})$$

The random variables $\xi_i^{(n)}$ counting the number of (derived) mutations in i copies (for $1 \leq i < n$) can be expressed in terms of $L_{k,\ell}^{(n)}$ as follows. Write $M_{k,\ell}^{(n)}$ for the number of mutations that occur on the ℓ th edge while there are k blocks in our n - Λ -coalescent, with $M_{k,\ell}^{(n)} \equiv 0$ if k blocks are never realized. Then

$$\xi_i^{(n)} = \sum_{k=2}^n \sum_{\ell=1}^k \mathbf{1}_{(L_{k,\ell}^{(n)}=i)} M_{k,\ell}^{(n)}. \quad (\text{A16})$$

Proof. As in (A12), let $L_{k,\ell}^{(n)}$ denote the random number of leaves—of n —subtended to block ℓ when there are k blocks, with $L_{k,\ell}^{(n)} = 0$ if k blocks are never realized.

Distinguishing cases as before, we obtain

$$\begin{aligned} \mathbb{E}[\xi_i^{(n)} \xi_j^{(n)}] &= \sum_{k=2}^n \sum_{\ell=1}^k \sum_{k'=2}^n \sum_{\ell'=1}^{k'} \mathbb{P}\{L_{k,\ell}^{(n)} = i, L_{k',\ell'}^{(n)} = j\} \mathbb{E}[M_{k,\ell}^{(n)} M_{k',\ell'}^{(n)}] \\ &= \sum_{k=2}^n k(k-1) \mathbb{P}\{L_{k,1}^{(n)} = i, L_{k,2}^{(n)} = j\} \mathbb{E}[M_{k,1}^{(n)} M_{k,2}^{(n)}] \\ &\quad + \mathbf{1}_{(i=j)} \sum_{k=2}^n k \mathbb{P}\{L_{k,1}^{(n)} = i\} \mathbb{E}[(M_{k,1}^{(n)})^2] \\ &\quad + 2 \sum_{k=3}^n \sum_{k'=2}^{k-1} k k' \mathbb{P}\{L_{k,1}^{(n)} = i, L_{k',1}^{(n)} = j\} \mathbb{E}[M_{k,1}^{(n)} M_{k',1}^{(n)}]. \end{aligned}$$

The three expected values on the right-hand side can now be expressed in terms of the total jump rates of our block-counting process Y away from states k resp. k' . Indeed, for the first expectation on the rhs note that the time while there are k blocks is the same for both edges 1 and 2, but mutations are collected *independently*. Thus, the expected number of mutations can be computed as the second moment of an exponential random variable:

$$\begin{aligned}
\mathbb{E}[M_{k,1}^{(n)} M_{k,2}^{(n)}] &= \int_0^\infty \mathbb{E}[M_{k,1}^{(n)} M_{k,2}^{(n)} \mid k \text{ blocks for } t \text{ time units}] (-q_{kk}) e^{-(q_{kk})t} dt \\
&= \int_0^\infty \left(\mathbb{E}[M_{k,1}^{(n)} \mid k \text{ blocks for } t \text{ time units}] \right)^2 (-q_{kk}) e^{-(q_{kk})t} dt \\
&= \int_0^\infty \left(\frac{\theta}{2} t \right)^2 (-q_{kk}) e^{-(q_{kk})t} dt \\
&= \frac{\theta^2}{4} \frac{2}{(-q_{kk})^2}.
\end{aligned}$$

The second expectation is given by the second moment of a mixture of a Poisson random variable with an exponential distribution:

$$\begin{aligned}
\mathbb{E}[(M_{k,1}^{(n)})^2] &= \int_0^\infty \mathbb{E}[(M_{k,1}^{(n)})^2 \mid k \text{ blocks for } t \text{ time units}] (-q_{kk}) e^{-(q_{kk})t} dt \\
&= \int_0^\infty \left[\sum_{m=0}^\infty m^2 \frac{(\theta/2)t}{m!} e^{-(\theta/2)t} \right] (-q_{kk}) e^{-(q_{kk})t} dt \\
&= \int_0^\infty \left[\left(\frac{\theta}{2} t \right)^2 + \frac{\theta}{2} t \right] (-q_{kk}) e^{-(q_{kk})t} dt \\
&= \frac{\theta}{2} \frac{1}{-q_{kk}} + \frac{\theta^2}{4} \frac{2}{(-q_{kk})^2}.
\end{aligned}$$

The last expectation is (due to independence obtained from the memoryless property of coalescent jump times) just the product of the two individual expected values; *i.e.*,

$$\mathbb{E}[M_{k,1}^{(n)} M_{k',1}^{(n)}] = \frac{1}{(-q_{kk})(-q_{k'k'})}.$$

Altogether, we arrive at

$$\begin{aligned}
\mathbb{E}[\xi_i^{(n)} \xi_j^{(n)}] &= \frac{\theta^2}{4} \sum_{k=2}^n k(k-1) \mathbb{P}\{L_{k,1}^{(n)} = i, L_{k,2}^{(n)} = j\} \frac{2}{(-q_{kk})^2} \\
&\quad + \mathbf{1}_{(i=j)} \sum_{k=2}^n k \mathbb{P}\{L_{k,1}^{(n)} = i\} \left(\frac{\theta}{2} \frac{1}{-q_{kk}} + \frac{\theta^2}{4} \frac{2}{(-q_{kk})^2} \right) \\
&\quad + 2 \frac{\theta^2}{4} \sum_{k=3}^n \sum_{k'=2}^{k-1} k k' \mathbb{P}\{L_{k,1}^{(n)} = i, L_{k',1}^{(n)} = j\} \frac{1}{(-q_{kk})(-q_{k'k'})},
\end{aligned}$$

from which the result follows with the help of (A13), (A14), and (A15).

Let $B_i^{(n)}$ denote the total length of branches subtending i of n leaves. The covariance of $B_i^{(n)}$ and $B_j^{(n)}$ now follows immediately from previous results since

$$\begin{aligned}
\mathbb{E}[B_i^{(n)} B_j^{(n)}] &= \sum_{k=2}^n k(k-1) \mathbb{P}\{L_{k,1}^{(n)} = i, L_{k,2}^{(n)} = j\} \frac{2}{q_{kk}^2} \\
&\quad + \mathbf{1}_{(i=j)} \sum_{k=2}^n k \mathbb{P}\{L_{k,1}^{(n)} = i\} \frac{2}{q_{kk}^2} + \sum_{k=3}^n \sum_{k'=2}^{k-1} k k' \mathbb{P}\{L_{k,1}^{(n)} = i, L_{k',1}^{(n)} = j\} \frac{1}{q_{kk} q_{k'k'}}.
\end{aligned} \tag{A17}$$

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.156612/-/DC1>

Statistical Properties of the Site-Frequency Spectrum Associated with Λ -Coalescents

Matthias Birkner, Jochen Blath, and Bjarki Eldon

File S1

FU'S COVARIANCE FORMULAS FOR THE KINGMAN CASE

FU (1995) obtains expressions for the expected value $\mathbb{E}[\xi_i^{(n)}]$, variance $\text{Var}(\xi_i^{(n)})$, and covariance $\text{Cov}(\xi_i^{(n)}, \xi_j^{(n)})$ of the number $\xi_i^{(n)}$ of mutations in i copies in a sample of size n whose genealogy *is governed by the Kingman coalescent*. For ease of reference we now cite FU (1995)'s results. Let

$$a_n := 1 + \frac{1}{2} + \cdots + \frac{1}{n-1}$$

for $n \geq 2$ with $a_1 := 0$, and

$$\beta_n(i) := \frac{2n}{(n-i+1)(n-i)}(a_{n+1} - a_i) - \frac{2}{n-i}.$$

Define the quantities

$$\sigma_{ii} = \begin{cases} \beta_n(i+1), & \text{if } i < n/2, \\ 2\frac{a_n - a_i}{n-i} - i^{-2}, & \text{if } i = n/2, \\ \beta_n(i) - i^{-2}, & \text{if } i > n/2, \end{cases}$$

and, for $i < j$,

$$\sigma_{ij} = \begin{cases} \frac{\beta_n(j+1) - \beta_n(j)}{2}, & \text{if } i+j < n, \\ \frac{a_n - a_i}{n-i} + \frac{a_n - a_j}{n-j} - \frac{\beta_n(i+1) + \beta_n(j)}{2} - \frac{1}{ij}, & \text{if } i+j = n, \\ \frac{\beta_n(i) - \beta_n(i+1)}{2} - \frac{1}{ij}, & \text{if } i+j > n. \end{cases}$$

Theorem S1 (Fu 1995). *For the unfolded site frequency spectrum of a n -sample with mutation rate $\theta/2$ and genealogy given by a Kingman-coalescent,*

$$\begin{aligned} \mathbb{E}[\xi_i^{(n)}] &= \frac{\theta}{i}, \\ \text{Var}(\xi_i^{(n)}) &= \frac{\theta}{i} + \sigma_{ii}\theta^2, \\ \text{Cov}(\xi_i^{(n)}, \xi_j^{(n)}) &= \sigma_{ij}\theta^2, \end{aligned}$$

for $i, j = 1, \dots, n-1$ with $i \neq j$.

For small j and large n , the covariances turn negative, i.e. $\sigma_{ij} < 0$. Similar results hold for the folded site-frequency spectrum, where FU (1995) defines the folded spectrum, for a sample of size n , as

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

where $\delta_{i,n-i} = 1$ if $i = n - i$, and is zero otherwise.

TRANSITION RATES IN SPECIAL CASES

In this section we specify transition rates associated with Lambda-coalescents when the measure Λ assumes specific forms.

In the case when $\Lambda = \text{Beta}(a, b)$ ($a, b > 0$) with density

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} \mathbf{1}_{(0,1)}(x), \quad (1)$$

the q_{ij} can be computed a little more explicitly:

$$\begin{aligned} \lambda_{n,k} &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^{k+a-3} (1-x)^{n-k+b-1} dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(k+a-2)\Gamma(n-k+b)}{\Gamma(n-2+a+b)} \\ &= \frac{(a)_{k-2}(b)_{n-k}}{(a+b)_{n-2}}, \end{aligned}$$

where $(x)_i = x(x+1) \cdots (x+i-1)$, $(x)_0 = 1$, and we used $\Gamma(x+1) = x\Gamma(x)$. Thus

$$q_{ij} = \binom{i}{i-j+1} \lambda_{i,i-j+1} = \frac{i!}{(i-j+1)!(j-1)!} \frac{(a)_{i-j-1}(b)_{j-1}}{(a+b)_{i-2}}.$$

For coalescent (5), one obtains, for $2 \leq k \leq n$,

$$\begin{aligned} \lambda_{n,k} &= \int_0^1 x^{k+2} (1-x)^{n-k} \left(\frac{2}{2+\psi^2} \delta_0(dx) + \frac{\psi^2}{2+\psi^2} \delta_\psi(dx) \right) \\ &= \frac{2}{2+\psi^2} \mathbf{1}_{(k=2)} + \frac{\psi^2}{2+\psi^2} \left[\psi^{k+2} (1-\psi)^{n-k} \right]. \end{aligned}$$

In particular, for $\psi = 0$, one has $\lambda_{n,k} = \mathbf{1}_{(k=2)}$, and for $\psi = 1$, this gives $\lambda_{n,k} = \frac{1}{3} \mathbf{1}_{(k=n)}$.

Figure S1: The results of BERESTYCKI *et al.* (2013) (see Thm. 1 in main text) regarding convergence of the scaled unfolded frequency spectrum compared to simulations with no. of leaves $n = 10000$ and varying over α and θ as shown.

Figure S2: The results of BERESTYCKI *et al.* (2013) (see Thm. 1 in main text) regarding convergence of the scaled unfolded frequency spectrum compared to simulations with no. of leaves $n = 10000$ and varying over α and θ as shown.

Figure S3: The results of BERESTYCKI *et al.* (2013) (see Thm. 1 in main text) regarding convergence of the scaled unfolded frequency spectrum (grey bars) compared to simulations (white bars) varying over α as shown, with $\theta = 1$. Panels in left column are for sample size $n = 500$, panels in right column are for sample size $n = 1000$.

Figure S4: The folded freq. spectrum (white bars) of the data of SIGURGÍSLASON and ÁRNASON (2003) along with predictions of the Kingman coalescent (dark-grey bars), and the Beta($2 - \hat{\alpha}, \hat{\alpha}$)-coalescent (light-grey bars). The vertical lines represent the standard deviation; obtained for the Beta($2 - \hat{\alpha}, \hat{\alpha}$)-coalescent from 10^5 iterations. The class labelled '7' represents the collated tail of the spectrum, from 11 to $74/2$.

Figure S5: Let $L_i^{(n)}$ denote the random length of branches subtending i of n leaves, and let $L^{(n)}$ denote the total length of the genealogy. The ratio $p_i := \mathbb{E}[L_i^{(n)}] / \mathbb{E}[L^{(n)}]$ then corresponds to $\varphi_n(i)$. Define $R_i := L_i/L$, and denote by $\overline{R}_i$ an estimate of $\mathbb{E}[R_i]$, the correct quantity we approximate with $\varphi_n(b)$. Comparison of the ratios $p_i := \mathbb{E}[L_i]/\mathbb{E}[L]$ and $\overline{R}_i$ for the Beta($2 - \alpha, \alpha$)-coalescent with no. of leaves n and $\alpha = a$ varying as shown in the legend. Entries are $|p_i - \overline{R}_i| / \overline{R}_i$, in which $\overline{R}_i$ are estimates of $\mathbb{E}[R_i]$ obtained from 10^5 iterations.

Figure S6: Let $L_i^{(n)}$ denote the random length of branches subtending i of n leaves, and let $L^{(n)}$ denote the total length of the genealogy. The ratio $p_i := \mathbb{E}[L_i^{(n)}] / \mathbb{E}[L^{(n)}]$ then corresponds to $\varphi_n(i)$. Define $R_i := L_i/L$, and denote by $\overline{R}_i$ an estimate of $\mathbb{E}[R_i]$, the correct quantity we approximate with $\varphi_n(b)$. Comparison of the ratios $p_i := \mathbb{E}[L_i]/\mathbb{E}[L]$ and $\overline{R}_i$ for the Beta($2 - \alpha, \alpha$)-coalescent with no. of leaves n and $\alpha = a$ varying as shown in the legend. Entries are $|p_i - \overline{R}_i| / \overline{R}_i$, in which $\overline{R}_i$ are estimates of $\mathbb{E}[R_i]$ obtained from 10^5 iterations.

Approximate bayesian computation (ABC; PRITCHARD *et al.*, 1999; BEAUMONT *et al.*, 2002, 2009; MARIN *et al.*, 2011) seems well suited to distinguish Lambda-coalescent processes from the Kingman coalescent, and to obtain credible intervals for the coalescence parameters. By way of example, ŽIVKOVIĆ & TELLIER (2012) apply ABC on the frequency spectrum to distinguish between possible demographic scenarios in relation to seed banks. Although we do not apply ABC for now, we do consider the distribution of the ℓ^2 -norm when one compares the observed spectrum with one simulated from a Lambda-coalescent. Thus, let

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

denote simulated spectrum obtained under a particular coalescent process. The ℓ^2 -norm is now

$$\ell^2 = \sqrt{\sum_i \left(\xi_i^n - \xi_i^{(*,n)} \right)^2}$$

where $\xi_i^{(n)}$ denotes the observed spectrum. Figures S7–S9 in show the distribution of the ℓ^2 -norm for the different Atlantic cod datasets based on 1000 iterations for each parameter value. Each vertical line denotes the interval spanned by the quantiles 5% and 95%, while the closed square on each line denotes the median. By way of example, one would clearly reject the Kingman coalescent in favor of the beta-coalescent for the ÁRNASON (2004) dataset (Figure S11). In Figures S12 and S13 the expected values for the beta-coalescent for $\alpha = 1.5$ and $\alpha = 1.05$, resp., replaces the data; i.e. we consider the distribution of the ℓ^2 -norm if the data were exactly the expected site-frequency spectrum associated with the beta-coalescent. The Figures indicate that an analysis of the distribution of the ℓ^2 -norm is a promising way to distinguish Lambda-coalescents from the Kingman coalescent.

Figure S7: The 5% and 95% quantiles (lines) and the median (squares) of the ℓ^2 -distance between simulated data for each parameter value (α or ψ) and the Faroe Islands cod dataset of SIGURGÍSLASON and ÁRNASON (2003). The quantiles for each parameter value are based on 1000 iterations. The α value 1.435 yields the minimum 95% quantile of 17.177, while the value 0.06 of ψ yields the minimum 95% quantile of 19.081.

Figure S8: The 5% and 95% quantiles (lines) and the median (squares) of the ℓ^2 -distance between simulated data for each parameter value (α or ψ) and the Greenland cod dataset of ÁRNASON *et al.* (2000). The quantiles for each parameter value are based on 1000 iterations. The α value 1.665 yields the minimum 95% quantile of 5.661, while the value 0.025 of ψ yields the minimum 95% quantile of 5.839.

Figure S9: The 5% and 95% quantiles (lines) and the median (squares) of the ℓ^2 -distance between simulated data for each parameter value (α or ψ) and the Iceland cod dataset of ÁRNASON *et al.* (2000). The quantiles for each parameter value are based on 1000 iterations. The α value 1.405 yields the minimum 95% quantile of 9.168, while the value 0.02 of ψ yields the minimum 95% quantile of 9.806.

Figure S10: The 5% and 95% quantiles (lines) and the median (squares) of the ℓ^2 -distance between simulated data for each parameter value (α or ψ) and the joint Iceland and Greenland cod dataset of ÁRNASON *et al.* (2000). The quantiles for each parameter value are based on 1000 iterations. The α value 1.535 yields the minimum 95% quantile of 8.426, while the value 0.015 of ψ yields the minimum 95% quantile of 10.863.

Figure S11: The 5% and 95% quantiles (lines) and the median (squares) of the ℓ^2 -distance between simulated data for each parameter value (α or ψ) and the Atlantic cod dataset of ÁRNASON (2004). The quantiles for each parameter value are based on 1000 iterations. The α value 1.545 yields the minimum 95% quantile of 11.874, while the value 0.01 of ψ yields the minimum 95% quantile of 16.735.

Figure S12: The 5% and 95% quantiles (lines) and the median (squares) of the ℓ^2 -distance between simulated data for each parameter value (α or ψ) and expected values of the site-frequency spectrum associated with the Beta($2 - \alpha, \alpha$)-coalescent when $\alpha = 1.5$, sample size is $n = 1000$ and $\theta = 1.0$. The quantiles for each parameter value are based on 1000 iterations. The α value 1.505 yields the minimum 95% quantile of 8.035, while the value 0.01 of ψ yields the minimum 95% quantile of 10.587.

Figure S13: The 5% and 95% quantiles (lines) and the median (squares) of the ℓ^2 -distance between simulated data for each parameter value (α or ψ) and expected values of the site-frequency spectrum associated with the Beta($2 - \alpha, \alpha$)-coalescent when $\alpha = 1.05$, sample size is $n = 1000$ and $\theta = 1.0$. The quantiles for each parameter value are based on 1000 iterations. The α value 1.205 yields the minimum 95% quantile of 28.472, while the value 0.04 of ψ yields the minimum 95% quantile of 26.318.

Figure S14: Estimates ($\hat{\alpha}$) obtained by applying the pseudo-likelihood to each of 1000 simulated datasets for sample size $n = 100$ obtained for value of $\alpha = a$ and θ as shown. Solid vertical lines indicate the true values of $\alpha = a$, and the broken lines indicate the 5%, 50%, and 95% quantiles.

Figure S15: Estimates ($\hat{\psi}$) obtained by applying the pseudo-likelihood to each of 1000 simulated datasets for sample size $n = 100$ obtained for value of ψ and θ as shown. Solid vertical lines indicate the true values of ψ , and the broken lines indicate the 5%, 50%, and 95% quantiles.

Figure S16: Estimates ($\hat{\psi}$ and $\hat{\alpha}$) obtained by applying the pseudo-likelihood to each of 1000 simulated datasets for sample size $n = 100$ under the Kingman coalescent and varying over θ as shown. Broken lines indicate the 5%, 50%, and 95% quantiles.

References

- ÁRNASON, E., 2004 Mitochondrial cytochrome *b* variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. *Genetics* 166: 1871–1885.
- ÁRNASON, E., P. H. PETERSEN, K. KRISTINSSON, H. SIGURGÍSLASON, and S. PÁLSSON, 2000 Mitochondrial cytochrome *b* DNA sequence variation of Atlantic cod from Iceland and Greenland. *J. Fish Biol* 56: 409–430.
- BEAUMONT, M., W. ZHANG, and D. BALDING, 2002 Approximate bayesian computation in population genetics. *Genetics* 162: 2025–2035.
- BEAUMONT, M. A., J. CORNUET, J. MARIN, and C. P. ROBERT, 2009 Adaptive approximate bayesian computation. *Biometrika* 96: 983–990.
- BERESTYCKI, J., N. BERESTYCKI, and V. LIMIC, 2013 A sampling formulae for lambda-coalescents. *ArXiv:1201.6512*.
- FU, Y., 1995 Statistical properties of segregating sites. *Theor Popul Biol* 48: 172–197.
- MARIN, J., P. PUDLO, C. ROBERT, and R. RYDER, 2011 Approximate bayesian computational methods. *Statistics and Computing* 22: 1167–1180.
- PRITCHARD, J. K., M. T. SEIELSTAD, A. PEREZ-LEZAUN, and M. W. FELDMAN, 1999 Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molec Biol Evol* 16: 1791–1798.
- SIGURGÍSLASON, H., and E. ÁRNASON, 2003 Extent of mitochondrial DNA sequence variation in Atlantic cod from the Faroe Islands: a resolution of gene genealogy. *Heredity* 91: 557–564.
- ŽIVKOVIĆ, D., TELLIER, A., 2012 Germ banks affect the inference of past demographic events. *Mol Ecol* 21: 5434–5446.