Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents?

Matthias Birkner[1], Jochen Blath[2], Bjarki Eldon[2,*], Fabian Freund[3]

August 6, 2014

Author affiliations:

1. JGU Mainz, Institut für Mathematik
55099 Mainz, Germany

2. TU Berlin, Institut für Mathematik
10623 Berlin, Germany

3. University of Hohenheim, Institute of plant breeding, seed science, and population genetics
70599 Stuttgart, Germany

*: corresponding author

Running title: Multiple mergers and population growth

Keywords: coalescent, multiple mergers, population growth, approximate bayesian computation, statistical power, Atlantic cod, site-frequency spectrum

corresponding author:
Bjarki Eldon
TU Berlin, Institut für Mathematik
Straße des 17. Juni 136
10623 Berlin, Germany
Email: eldon@math.tu-berlin.de
Phone: +49 303 1425 762
Fax: +49 +(0) 30 314 21695

Abstract

The ability of the site-frequency spectrum (SFS) to reflect the particularities of gene genealogies exhibiting multiple mergers of ancestral lines as opposed to those obtained in the presence of exponential population growth is our focus. An excess of singletons is a well-known characteristic of both population growth and multiple mergers. Other aspects of the SFS, in particular the weight of the right tail, are, however, affected in specific ways by the two model classes. Using minimum-distance statistics, and an approximate likelihood method, our estimates of statistical power indicate that exponential growth can indeed be distinguished from multiple merger coalescents, even for moderate sample size, if the number of segregating sites is high enough. Additionally, we use a normalised version of the SFS as a summary statistic in an approximate bayesian computation (ABC) approach to distinguish multiple mergers from exponential population growth. The ABC approach gives further positive evidence as to the general eligibility of the SFS to distinguish between the different histories, but also reveals that suitable weighing of parts of the SFS can improve the distinction ability. The important issue of the difference in timescales between different coalescent processes (and their implications for the scaling of mutation parameters) is also discussed.

Introduction

The site-frequency spectrum (SFS) at a given locus is one of the most important and popular statistics based on genetic data sampled from a natural population. In combination with the postulation of the assumptions of the infinitely-many sites mutation model (Watterson, 1975) and a suitable underlying coalescent framework, the SFS allows one to draw inference about evolutionary parameters, such as coalescent parameters associated with multiple-merger coalescents or population growth models.

The Kingman coalescent, developed by Kingman (1982a,b,c), Hudson (1983a,b) and Tajima (1983), describing the random ancestral relations among DNA sequences drawn from natural populations, is a prominent and widely-used coalescent model from which one can make predictions about genetic diversity. Many quantities of interest, such as the expected values and covariances of the SFS, are easily computed (Fu, 1995) from the Kingman coalescent. Its robustness is quite remarkable, and indeed a large number of models can be shown to have the Kingman coalescent, or a variant thereof, as their limit process, cf. e.g. Möhle (1998). A large volume of work is thus devoted to inference methods based on the Kingman coalescent; see e.g. Donnelly and Tavaré (1995), Hudson (1990), Nordborg (2001), Hein *et al.* (2005) or Wakeley (2007) for reviews.

However, many evolutionary histories can lead to significant deviations from the Kingman coalescent model. Such deviations can be detected using a variety of statistical tools, such as Tajima's D (Tajima, 1989a), Fu and Li's D (Fu and Li, 1993) or Fay and Wu's H (Fay and Wu, 2000), which are all functions of the SFS. However, they do not always allow to identify the actual evolutionary mechanisms leading to such deviations. Developing statistical tools that allow to distinguish between different evolutionary histories is, therefore, of fundamental importance.

The present work focuses on properties of the (folded and unfolded) SFS in the infinitely-many sites model for three population histories: (1) classical Kingman coalescent, (2) population growth, in particular exponential population growth, and (3) high fecundity coupled with skewed offspring distributions (HFSOD), resulting in gene genealogies being described

by so-called Lambda-coalescents (SAGITOV, 1999; PITMAN, 1999; DONNELLY and KURTZ, 1999). Briefly, multiple merger coalescents may be more appropriate for organisms exhibiting HFSOD than the Kingman coalescent (cf. eg. BECKENBACH, 1994; ÁRNASON, 2004; ELDON and WAKELEY, 2006; SARGSYAN and WAKELEY, 2008; HEDGECOCK and PUDOVKIN, 2011), see also a recent review by TELLIER and LEMAIRE (2014).

Both recent population growth as well as multiple-merger coalescents may lead to an excess of singletons in the SFS compared to the classical Kingman coalescent based SFS, which e.g. contributes to shifting Tajima's D values to the negative. Indeed, DURRETT and SCHWEINSBERG (2005) prove that TAJIMA (1989b)'s D will be negative, at least for large sample size, under fairly general multiple-merger coalescents.

The associated genealogical trees are, however, qualitatively different. While moderate fluctuations in population size lead to a time-change of the Kingman coalescent (KAJ and KRONE, 2003), multiple merger coalescents by definition change the topology of the genealogical tree. There is thus hope that each demographic effect leaves specific signatures in the resulting SFS, not only with respect to an excess of singletons, but for example also with respect to its right tail.

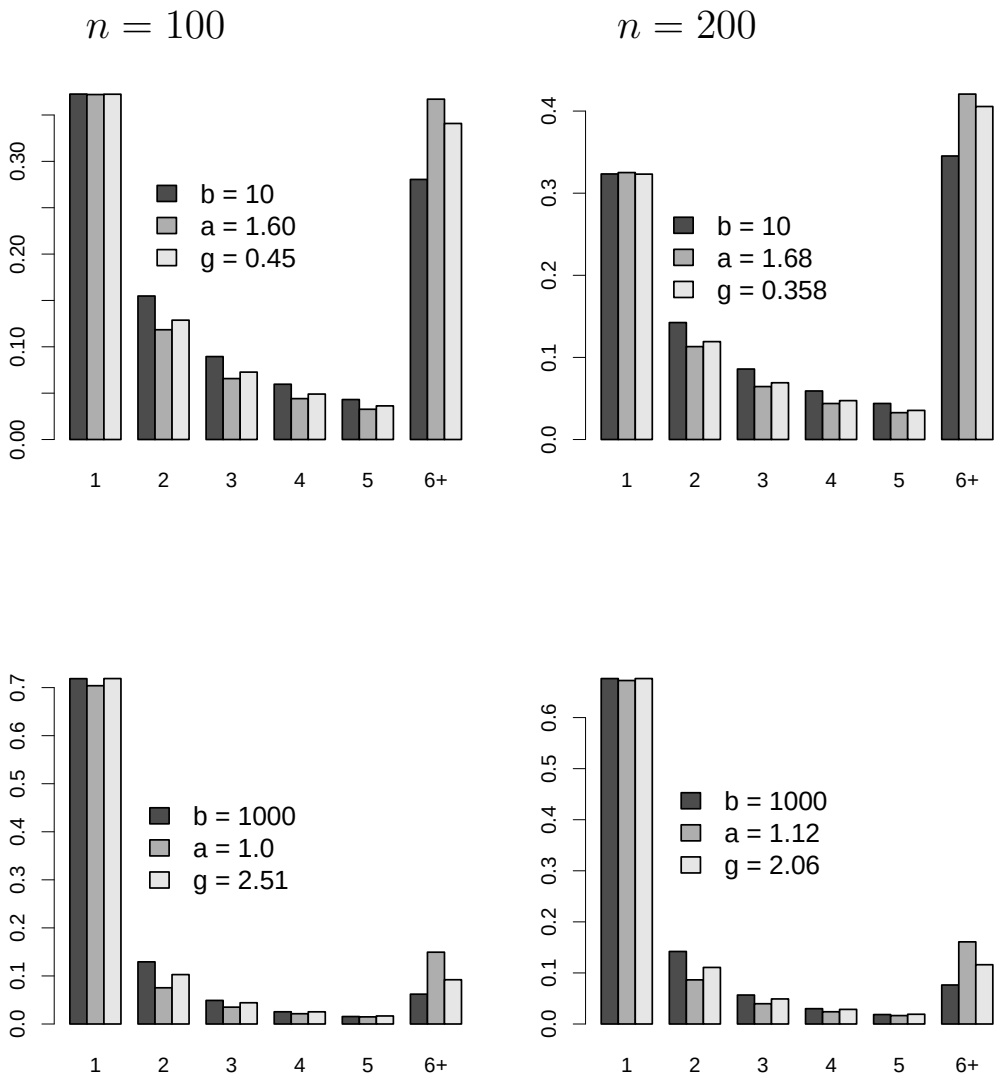
Indeed, one observes that the Kingman coalescent will not be a good match to genetic data containing many singleton polymorphisms due to a lack of free (coalescent) parameters, as opposed to multiple merger and population growth models, which both can predict an excess of singletons. Encouragingly, multiple merger and growth models exhibit noticeable differences in the bulk of the site-frequency spectrum, in particular in the lumped tail (Figure 1). In Figure 1, the normalised expected spectrum $\varphi_i^{(n),\Pi}$ (2) for a given coalescent Π, ie. the expected spectrum scaled by the expected total tree length, is compared for a particular multiple-merger coalescent (B; SCHWEINSBERG, 2003), and exponential (E) growth models for sample size (number of leaves) n as shown. The first five classes (representing relative length of external branches, two-leaf branches, etc.) are shown, with classes from six onwards collected together (labelled as '6+'). In Figure 1, the relative external branch lengths were matched between the different coalescent processes. Even though the relative exter-

nal branch lengths, and by implication the relative number of singletons, can be matched between the different processes, the lumped tail (group 6+ in Figure 1) differs between the multiple-merger coalescent (B), and exponential growth (E).

Matching the relative external branch lengths $\varphi_1^{(n),\Pi}$ (2) and observing how the rest of the normalised spectrum behaves, as illustrated in Figure 1, gives hope that multiple merger processes may be distinguished from (at least) particular population growth models with adequate statistical power. The quantity $\varphi_1^{(n),\Pi}$ varies as a function of n and the associated coalescent parameter (Figure S1). In the limit of large n, for the Kingman coalescent, $\varphi_1^{(n),\mathrm{K}} = O(1/\log(n))$.

Inference methods for distinguishing population growth from the usual Kingman coalescent have been extensively studied, see e.g. Tajima (1989a), Slatkin and Hudson (1991), Rogers and Harpending (1992), Kaj and Krone (2003), Ramos-Onsins and Rozas (2002), and Sano and Tachida (2005). Detecting multiple merger coalescents in populations deviating from the Kingman coalescent assumptions is a relatively new direction of research. Indeed, deriving inference methods based on multiple merger coalescents has only just begun (Eldon and Wakeley, 2006; Birkner and Blath, 2008; Eldon, 2011; Birkner *et al.*, 2011, 2013a,b; Steinrücken *et al.*, 2013; Koskela *et al.*, 2013). In particular, Birkner *et al.* (2013b) obtain recursions for the expected site-frequency spectrum associated with Lambda-coalescents. In the present work we address the issue of detecting multiple merger coalescents from exponential population growth, using methods based on the SFS, by estimating statistical power for both point and interval hypotheses. As an alternative approach, we also consider a simple implementation of approximate Bayesian computation (ABC; Rubin, 1984; Tavaré *et al.*, 1997; Pritchard *et al.*, 1999; Cucala and Marin, 2013; Baragatti and Pudlo, 2014).

Figure 1: Matching $\varphi_1^{(n),\Pi}$ (2) for the different coalescent processes $\Pi \in \{\mathtt{E}, \mathtt{B}, \mathtt{A}\}$ for β (b), α (a), and γ (g) for algebraic growth (see Supporting Information) for comparison, and number of leaves n as shown. Expected values were computed exactly.



Theory and Methods

Basic properties of the site-frequency spectrum

Consider a sample of n DNA sequences taken at a given genetic locus and assume that we can distinguish between derived (new mutations) and ancestral states. For $n \in \mathbb{N}$ let $[n] := \{1, \ldots, n\}$. We denote by $\xi_i^{(n)}$ the total number of sites at which the mutant base appears $i \in [n-1]$ times. Then,

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

is referred to as the *unfolded* site-frequency spectrum based on the n DNA sequences. If mutant and wild-type cannot be distinguished, one often considers the *folded* spectrum $\underline{\eta}^{(n)} := \left(\eta_1^{(n)}, \ldots, \eta_{\lfloor n/2 \rfloor}^{(n)}\right)$, where ancestral and derived states are not distinguished, and hence

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

(Fu, 1995). In this study, we will mostly be concerned with the unfolded site-frequency spectrum. Define $\zeta_i^{(n)} := \xi_i^{(n)} / |\xi^{(n)}|$ where $|\xi^{(n)}| := \xi_1^{(n)} + \cdots + \xi_{n-1}^{(n)}$ denotes the total number of segregating sites. Thus, $\underline{\zeta}^{(n)} = \left(\zeta_1^{(n)}, \ldots, \zeta_{n-1}^{(n)}\right)$ is the 'normalized' unfolded SFS, with the convention that $\underline{\zeta}^{(n)} = 0$ in the trivial case of complete absence of segregating sites ($|\xi^{(n)}| = 0$).

In order to compute expected values, variances and covariances of the SFS, an explicit underlying probabilistic model is needed. In the following we assume that the genealogy of a sample can be described by a coalescent process, more precisely by either (a timechange of) the Kingman coalescent or a multiple-merger coalescent. In addition, the infinitely-many-sites mutation model (Watterson, 1975) is assumed, and mutations are modeled by a Poisson-process on the coalescent branches with rate $\theta/2$.

Closed-form expressions for the expected values and (co)variances of $\underline{\xi}^{(n)}$ have been determined in Fu (1995) when associated with the Kingman coalescent. One can represent the

expected values of $\underline{\xi}^{(n)}$ in a unified way using the results of GRIFFITHS and TAVARÉ (1998), KAJ and KRONE (2003) and BIRKNER *et al.* (2013b), that allow to treat the expected values (and covariances) of the SFS for all coalescent models in questions.

Let $\Pi^n = (\Pi^n_t, t \geq 0)$ be a (partition-valued exchangeable) coalescent process started from n leaves (partition blocks) corresponding to the random genealogy of a sample of size n. By discussing 'leaves' rather than DNA sequences we are emphasizing our viewpoint of the genealogy as a random graph, where the leaves are a particular kind of vertices. Our emphasis is on the topology of the genealogy, rather than the associated site-frequency spectrum.

If the initial number of leaves is not specified, we simply speak of Π. One may think of Π as the Kingman coalescent, but the point is that the following result will stay true also for externally time-changed Kingman coalescents as well as asynchronous multiple merger coalescents (a.k.a. 'Lambda'-coalescents in the mathematical literature), and even externally time-changed multiple merger coalescents.

Given n and a coalescent model Π, let $(Y^{(n)}_t)_{t\geq 0}$ be the block counting process of the underlying coalescent Π^n started from n lineages, i.e. $Y^{(n)}_t$ gives the number of ancestral lines (blocks) present/active at (backwards) time t. For $2 \leq k < n$, let $T^{(n)}_k$ be the random amount of time that $Y^{(n)}$ spends in state k. Given a coalescent Π^n started from n (unlabelled) lineages, denote by $p^{(n),\Pi}[k,i]$ the probability that, *conditional* on $Y^{(n)}$ taking value k, a given one of the k blocks subtends exactly $i \in [n-1]$ leaves. A general representation of $\mathbb{E}^{\Pi}\left[\xi^{(n)}_i\right]$ is then

$$\mathbb{E}^{\Pi}\left[\xi^{(n)}_i\right] = \frac{\theta}{2} \sum_{k=2}^{n-i+1} p^{(n),\Pi}[k,i] \cdot k \cdot \mathbb{E}^{\Pi}\left[T^{(n)}_k\right], \tag{1}$$

and for the normalized expected SFS $\varphi^{(n),\Pi}_i$

$$\varphi^{(n),\Pi}_i = \frac{\sum_{k=2}^{n-i+1} p^{(n),\Pi}[k,i] \cdot k \cdot \mathbb{E}^{\Pi}\left[T^{(n)}_k\right]}{\sum_{\ell=2}^{n} \ell\, \mathbb{E}^{\Pi}\left[T^{(n)}_\ell\right]}. \tag{2}$$

One can interpret the quantity $\varphi^{(n),\Pi}_i$ as the probability that a mutation, under the infinitely

many sites assumption and the coalescent model Π, with known ancestral types, appears i times in a sample of size n. Importantly, $\varphi_i^{(n),\Pi}$ is not a function of the mutation rate, unlike $\mathbb{E}^{\Pi}\left[\xi_i^{(n)}\right]$. One can also view $\varphi_i^{(n),\Pi}$ as a first-order approximation of the expected value $\mathbb{E}^{\Pi}\left[\zeta_i^{(n)}\right]$ of the normalised SFS.

As examples for Π we will consider the classical Kingman coalescent (`K`), exponential growth (`E`), and the Beta$(2-\alpha,\alpha)$ multiple-merger coalescent (`B`). These are recalled in the Supporting Information. The quantity $\varphi_i^{(n),\Pi}$ can be compared with $\zeta_i^{(n)}$ in a minimum-distance statistic (see below). Simulations suggest that $\varphi_i^{(n),\mathsf{B}}$ is a decent approximation of $\mathbb{E}^{\mathsf{B}}\left[\zeta_i^{(n)}\right]$ when α is not too close to 1, and n not too small (Birkner *et al.*, 2013b). Similar conclusions hold in the case of exponential growth (Figure S2).

Recursive formulae for the covariances of the SFS can also be established; see e.g. (23) in Fu (1995) for the Kingman case or Birkner *et al.* (2013b) for the Lambda-case. For general models of changes in population size cf. section 'The SFS under variable population size' in the Supporting Information.

Comparing the observed $\zeta_i^{(n)}$ (instead of $\xi_i^{(n)}$) to its expected value $\mathbb{E}^{\Pi}\left[\zeta_i^{(n)}\right]$ – obtained under a particular coalescent model Π – enables one to do inference without having to jointly estimate the mutation rate θ using e.g. a minimum-distance statistic. Unfortunately, there seems to be no explicit way of representing $\mathbb{E}^{\Pi}\left[\zeta_i^{(n)}\right]$ as a simple function of the coalescent parameters and sample size n. One may, instead, compare $\zeta_i^{(n)}$ to (2) (Birkner *et al.*, 2013b).

Timescales, segregating sites and mutation rates

The choice of a coalescent model (resp. demographic history) Π and its underlying parameters strongly affects classical estimates for the coalescent mutation rate $\theta/2$ (i.e. the Poisson rate at which mutations appear on coalescent branches). Assume w.l.o.g. for all multiple merger coalescents in question that the underlying coalescent measure Λ is always a probability measure: This normalisation fixes the coalescent time unit as the expected time to the most recent common ancestor of two individuals sampled uniformly from the population. In

particular, θ can then be interpreted as the expected number of observed pairwise differences in a sample of size two.

Given an observed number of segregating sites S in a sample of size n, a common estimate $\hat{\theta}^{\Pi}$ of the scaled mutation rate θ is the Watterson estimate

$$\hat{\theta}^{\Pi} := \frac{2S}{\mathbb{E}^{\Pi}\left[B^{(n)}\right]}, \tag{3}$$

where $\mathbb{E}^{\Pi}\left[B^{(n)}\right]$ is the expected total tree length of the underlying coalescent model Π. One can of course also estimate θ as a linear combination of the site-frequency spectrum (cf. Achaz, 2009) in the case of the Kingman coalescent. Using the recursions for $\mathbb{E}^{\Pi}\left[\xi_i^{(n)}\right]$ obtained by Birkner *et al.* (2013b), one can now also estimate θ analogously in case of a Lambda-coalescent.

The real-time embedding of genealogies can vary drastically between different model classes. Mathematically, coalescent processes can be obtained as the limits of genealogies in Cannings models (Cannings, 1974, 1975) (resp. similar models) under a suitable time-change. The real-time embedding is the expression of coalescent times on the time scale of the underlying Cannings models. This important aspect of coalescent models is subtly hidden in the actual resulting limiting models. For example, given a Cannings population model of fixed size N, let c_N be the probability that two gene copies, drawn uniformly at random and without replacement from a population of size N, derive from a common parental gene copy in the previous generation (see (S10) in the Supporting Information for an explicit formula and further details). While for the usual Wright-Fisher haploid model $c_N = 1/N$, in a population model studied by Schweinsberg (2003), which leads to the Beta$(2-\alpha, \alpha)$-coalescent, c_N is proportional to $1/N^{\alpha-1}$, for $1 < \alpha \leq 2$. By a limit theorem for Cannings models of Möhle and Sagitov (2001), one coalescent time unit corresponds to $1/c_N$ generations in the original model with population size N. In this framework, the expected total tree length $\mathbb{E}^{\mathsf{B}}[B^{(n)}]$ (measured in coalescent time units) *decreases* as a function of $\alpha \in (1, 2]$, while the corresponding quantity (measured in generations) $\mathbb{E}^{\mathsf{B}}[B^{(n)}]/c_N$ *increases* (Figure S3).

Accordingly, the mutation rate μ at the locus under consideration per individual per

generation must thus be scaled with $1/c_N$ (as noted e.g. in ELDON and WAKELEY (2006)), and the relation between μ, the coalescent mutation rate $\theta/2$ and c_N is then given by the (approximate) identity

$$\frac{\theta}{2} = \frac{\mu}{c_N}. \tag{4}$$

This allows one to obtain an approximate real-time calibration of the coalescent time unit $1/c_N$, given external knowledge of the per-generation mutation rate μ, and an estimate for θ, e.g. based on (3).

The time-scaling applied to a classical Wright-Fisher model with *fluctuating* population size (as in KAJ and KRONE (2003)) in order to obtain a (time-changed) Kingman coalescent is recalled in the Supporting Information, see in particular Equation (S11). Again, the estimate (3) of θ depends on the exponential growth parameter β.

Our aim is to construct a test resp. a decision rule to distinguish between the two model classes `E` and `B` (which intersect exactly in `K`). Hence, our methods can also be used to distinguish growth from stationary models and multiple merger coalescents from the Kingman coalescent model, thus complementing existing literature (cf. eg. RAMÍREZ-SORIANO *et al.*, 2008; RAMOS-ONSINS and ROZAS, 2002). Our investigation can be applied to more general Lambda-coalescents or other growth models. In the Supporting Information, we additionally consider the case of *algebraic* (power law) population growth, which may be applicable when the geometry of a habitat (say, a coastline) restricts the growth of a population.

Approximate likelihood ratio tests for the SFS

In order to distinguish, say, `E` from `B` based on an observed site-frequency spectrum, a natural approach is to construct a (unnested) likelihood-ratio hypothesis test. Suppose our null-hypothesis H_0 is presence of recent exponential population growth `E` with some parameter $\beta \in [0, \infty)$, and we wish to test it against the alternative H_1 hypothesis of a multiple merger coalescent, say, the Beta$(2-\alpha, \alpha)$-coalescent (SCHWEINSBERG, 2003) for some $\alpha \in [1, 2]$.

To clarify our notation, define the hypotheses we will work with;

$$\Theta_0 := \{(\text{exponential growth } (\beta), \theta) : \beta \in [0, \infty), \quad \theta \in (0, \infty)\}$$

and

$$\Theta_1 := \{(\text{Beta}(2-\alpha, \alpha)\text{-coalescent}, \theta) : \alpha \in [1, 2], \quad \theta \in (0, \infty)\}$$

where we interpret $\beta = 0$ and $\alpha = 2$ as corresponding to the Kingman coalescent. Later, we will fix the expected total number of segregating sites, which will make θ a function of β (resp. α). As shorthand, we will use $\Theta_\beta := [0, \infty)$ and $\Theta_\alpha = [1, 2]$ where the intervals refer to β resp. α. Recall that $\underline{\xi}^{(n)}$ is the observed site frequency spectrum for a sample of size n and, given mutation rate $\theta/2 > 0$, let $\varphi^{(n),\Pi}$ (2) be the normalized expected site frequency spectrum.

Let $H_0 := \{\vartheta \in \Theta_0\}$ denote the null-hypothesis and $H_1 = \{\vartheta \in \Theta_1\}$ denote the alternative. Let $L(\vartheta, \underline{\xi}^{(n)})$ be the likelihood function of the observed frequency spectrum under model ϑ. Then, one can define the likelihood-ratio function

$$\varrho(\underline{\xi}^{(n)}) := \frac{\sup\{L(\vartheta, \underline{\xi}^{(n)}), \vartheta \in \Theta_0\}}{\sup\{L(\vartheta, \underline{\xi}^{(n)}), \vartheta \in \Theta_1\}}. \tag{5}$$

One would reject the null-hypothesis H_0 if ϱ is sufficiently small. More precisely, given a significance level $a \in (0, 1)$ (say, $a = 0.05$), one needs to determine a constant ϱ^* such that

$$\sup_{\vartheta \in \Theta_0} \mathbb{P}_\vartheta \Big\{ \varrho(\underline{\xi}^{(n)}) \leq \varrho^* \Big\} \leq a. \tag{6}$$

The corresponding power function G is then given by

$$G(\vartheta) = \mathbb{P}_\vartheta\{\varrho(\underline{\xi}^{(n)}) \leq \varrho^*\}, \quad \vartheta \in \Theta_1. \tag{7}$$

Exact likelihoods can be computed recursively or via simulation (see e.g. Eldon and Wakeley (2006) and Sargsyan and Wakeley (2008)) from the representation (8) where $B_i^{(n)}$

is the random length of branches subtending $i \in [n-1]$ leaves (DNA sequences) with $B^{(n)} := B_1^{(n)} + \cdots + B_{n-1}^{(n)}$ being the total length of the tree (recall $\mathbb{E}^\vartheta[\xi_i^{(n)}] = (\theta/2)\mathbb{E}^\vartheta[B_i^{(n)}]$) with expectation being taken relative to coalescent process Π with parameter ϑ (to ease notation, write $B_i \equiv B_i^{(n)}$),

$$\mathbb{P}^\vartheta\{\xi_i^{(n)} = k_i, i = 1, \ldots, n-1\} = \left(\frac{\theta}{2}\right)^{k_1+\cdots+k_{n-1}} \mathbb{E}^\vartheta\Big[\prod_{i=1}^{n-1} e^{-B_i\theta/2}\frac{B_i^{k_i}}{k_i!}\Big]. \tag{8}$$

For medium to large n, however, it is more practical to consider the approximate likelihood function

$$\widetilde{L}(\vartheta, \underline{\xi}^{(n)}) = \prod_{i=1}^{n-1} e^{-s\varphi_i^{(n)}} \frac{(\varphi_i^{(n)} s)^{k_i}}{k_i!}, \tag{9}$$

(where $s = \frac{\theta}{2}\mathbb{E}^\vartheta[B^{(n)}]$) pretending that the classes are approximately independent and Poisson-distributed (which is encouraged by the fact that the off-diagonal entries of the covariance matrix of $\underline{\xi}^{(n)}$ are small compared to the diagonal terms, see Birkner *et al.* (2013b)). The corresponding approximate likelihood-ratio will be denoted by $\widetilde{\varrho}$, and can then be used to determine quantiles associated with significance level a as in (6), and power function $\widetilde{G}$.

One often observes $\xi_i^{(n)} = 0$ for most i greater than some (small) number m in observed data, in particular for large n. It thus seems natural to consider (approximate) likelihood functions for "lumped spectra" (e.g. collapsing all entries in classes to the right of some number m into one class m^+).

Another natural type of lumping may be to collect together classes so that $\sum_i \varphi_i^{(n)} \geq x$ for some $x \in (0, 1/2]$. This may not always be feasible, though, if the individual $\varphi_i^{(n)}$ quickly become quite small, and we will refrain from going into a more detailed theoretical discussion of optimal lumpings.

Instead of (approximate) likelihoods, one can also consider rejection rules based on minimal distance statistics:

$$\varrho^{(d)} := \frac{\inf\{d(\underline{\varphi}^{(n)}(\vartheta), \underline{\xi}^{(n)}), \vartheta \in \Theta_0\}}{\inf\{d(\underline{\varphi}^{(n)}(\vartheta), \underline{\xi}^{(n)}), \vartheta \in \Theta_1\}}, \tag{10}$$

for some suitable distance measure d (e.g. the ℓ_p distance with $p = 2$) with corresponding power function $G^{(d)}$. If entries in the SFS become sparse and far apart, lumping the right tail should increase its relative contribution to the ℓ_2 distance, and might thus be more adequate for our purposes. In fact, we will observe such an effect in our subsequent analysis.

Additional information about the parameters in each model, for example reducing the respective parameter ranges to one-point hypotheses of type $H_0 = \{\vartheta = \beta^*\}$ vs. $H_1 = \{\vartheta = \alpha^*\}$, is expected to lead to substantially more powerful tests, in particular if both hypotheses are well-separated from the Kingman coalescent.

Approximate Bayes factors and model selection

Using the previous notation, an analogous Bayesian approach is to choose a model class based on a *Bayes factor*

$$\varrho^{\text{Bayes}} := \frac{\int_{\Theta_0} L(\vartheta, \underline{\zeta}^{(n)})\, d\pi_0(\vartheta)}{\int_{\Theta_1} L(\vartheta, \underline{\zeta}^{(n)})\, d\pi_1(\vartheta)},$$

given a pair of priors π_0, π_1 on Θ_0, Θ_1 (given the same prior probability of each model class, this is also the odds-ratio). We use the less informative normalized site frequency spectrum (nSFS) of an n sample, denoted by $\underline{\zeta}^{(n)}$, since it is robust to changes in mutation rates. Bayes factors based on (lumped) distances d and/or the folded nSFS may also be considered. In line with classical Bayes factor philosophy (cf. e.g. Kass and Raftery (1995)), one interprets an observed value of $\varrho^{\mathcal{B}} \gg 1$ as evidence in favor of Θ_0 over Θ_1. As exact likelihoods $L(\vartheta, \underline{\zeta}^{(n)})$ are often impractical, we employ approximate Bayesian methods (see e.g. Beaumont (2010)).

For the ABC analyses, we again focus on exponential growth (`E`) versus Beta coalescents (`B`) and denote the corresponding Bayes factor by $\varrho^{\mathtt{E/B}}$. Recall that for fixed n, both model classes can be parametrised by two parameters, mutation rate $\theta \in (0, \infty)$ and exponential growth rate $\beta \in [0, \infty)$ resp. the Beta coalescent parameter $\alpha \in [1, 2]$. To choose prior distributions on these two-dimensional parameter sets Θ_0 and Θ_1 we first record/assume the number s of segregating sites in the data. Then, we set marginal prior distributions for the

growth parameter β resp. the Beta coalescent parameter α. Finally, for each β resp. α, the mutation rate is determined as the Watterson estimate $\hat{\theta}^{\Pi}$ (3), which of course depends on β resp. α through the total branch length $\mathbb{E}^{\Pi}\left[B^{(n)}\right]$. We chose the mutation rate so that the expected number of mutations under the chosen coalescent model equals the Watterson estimate based on the (assumed) number of observed mutations.

For convenience, we employ a simple rejection-based ABC scheme to approximate the Bayes factor for the model (class) comparison given an observed nSFS (resp. folded and/or lumped versions, which can be treated analogously). First we simulate n_{reps} independent samples of the nSFS under each model class and record the distance to the observed nSFS in question. We then fix a tolerance level $x \in (0, 1)$ and count the number of simulations $n_{\mathtt{E}}$ from the growth model resp. $n_{\mathtt{B}}$ from the Beta coalescent model class that are among the $x\%$ best fits with respect to the ℓ_2 distance to the observed nSFS (the 'accepted' simulations). Here, we use an additional scaling by dividing each class (lumped classes) in the nSFS by the median (if non-zero) within this class observed in all simulations as implemented in the `R` package abc (Csillery *et al.*, 2012). The Bayes factor can then be approximated by

$$\varrho^{\mathtt{E/B}} \approx \frac{n_{\mathtt{E}}}{n_{\mathtt{B}}}.$$

To assess how well our ABC approach allows us to distinguish the model classes, we use two approaches from the `R` package abc. Both are based on leave-one-out cross-validation. More precisely, we pick n_{cv} simulations at random from each model class, treat them as the observed value of the nSFS and then run the ABC approach with the same parameters and simulations as above. For each cross-validation $i_M \in \{1, \ldots, n_{cv}\}$, with M denoting the model class `E` resp. `B`, we record the counts of accepted simulations $n_{\mathtt{E}}(i_M)$ and $n_{\mathtt{B}}(i_M)$ from each model class. As measures for the distinction ability of this approach, we record for each model class the (estimated) mean posterior probabilities π given the observed nSFS,

borrowing notation from STOEHR *et al.* (2014),

$$\mathbb{E}^{\mathtt{B}}\left[\pi(\mathtt{E}|\underline{\zeta}^{(n)})\right] \approx \frac{1}{n_{cv}} \sum_{i_{\mathtt{B}}=1}^{n_{cv}} \frac{n_{\mathtt{E}}(i_{\mathtt{B}})}{n_{\mathtt{E}}(i_{\mathtt{B}}) + n_{\mathtt{B}}(i_{\mathtt{B}})}$$

and

$$\mathbb{E}^{\mathtt{E}}\left[\pi(\mathtt{B}|\underline{\zeta}^{(n)})\right] \approx \frac{1}{n_{cv}} \sum_{i_{\mathtt{E}}=1}^{n_{cv}} \frac{n_{\mathtt{B}}(i_{\mathtt{E}})}{n_{\mathtt{E}}(i_{\mathtt{E}}) + n_{\mathtt{B}}(i_{\mathtt{E}})}$$

The mean misclassification probabilities are estimated as

$$\mathbb{E}^{\mathtt{B}}\left[\pi(\varrho^{\mathtt{E/B}} > 1|\underline{\zeta}^{(n)})\right] \approx \frac{1}{n_{cv}} \sum_{i_{\mathtt{B}}=1}^{n_{cv}} 1_{\{n_{\mathtt{E}}(i_{\mathtt{B}}) > n_{\mathtt{B}}(i_{\mathtt{B}})\}}$$

and

$$\mathbb{E}^{\mathtt{E}}\left[\pi(\varrho^{\mathtt{E/B}} < 1|\underline{\zeta}^{(n)})\right] \approx \frac{1}{n_{cv}} \sum_{i_{\mathtt{E}}=1}^{n_{cv}} 1_{\{n_{\mathtt{E}}(i_{\mathtt{E}}) < n_{\mathtt{B}}(i_{\mathtt{E}})\}}.$$

To ease notation we will omit n in the formulae.

In practice, we need to efficiently generate samples of the nSFS under the different models which can be achieved by backward-in-time coalescent simulations. For the exponential growth models (`E`), we use Hudson's ms (HUDSON (2002)) as implemented in the `R` (R CORE TEAM, 2012) package phyclust (CHEN (2011)). For the Beta-coalescents (`B`), we use custom `R` and `C` scripts to generate samples of the nSFS. To conduct the actual ABC analysis including cross-validation techniques, we employed the `R` package abc (CSILLÉRY *et al.* (2012)).

Results

Power estimates

To assess the sensitivity of our approximate likelihood ratio test associated with the likelihood ratio function (5), we consider its power G from (7) as a function of α with $H_0 = \Theta_\beta$ and $H_1 = \Theta_\alpha$. As shown in Figure 2, high power (at least 60%) to distinguish Beta(2 −

$\alpha, \alpha)$-coalescent from exponential growth can be obtained, in particular in the presence of sufficiently many segregating sites (Figure 2**B**). Similar conclusions hold for a smaller sample size ($n = 100$; Figures S12, S11). The number 300 of segregating sites is nearly the total number of polymorphisms (298) observed by CARR and MARSHALL (2008) who scanned whole mitochondrial genomes (15,655 bp) of the highly fecund Atlantic cod (*Gadus morhua*). For comparison, Figure 3 shows power estimates for observed parameters ($(n, s) = (30, 300)$) resembling the data considered by CARR and MARSHALL (2008). Reversing the hypotheses shows similarly promising power estimation results (Figure 4) for $n = 200$. We do not have a clear explanation for why the power is not monotone as a function of β, in particular for smaller Type I error.

When sample size is large, one will typically encounter many zeroes in a given observed site-frequency spectrum, so that we will lump the right tail of the spectrum at some threshold m^+. Our power estimates suggest that keeping at least the first five classes intact, and collecting the rest into one other class, has little effect on the power of the test (results not shown). Keeping only the singleton ($\xi_1^{(n)}$) class intact, and collecting all the rest into one class, however, significantly diminishes power (results not shown).

C (cf. KERNIGHAN and RITCHIE, 1988) code written for estimating the power, where use was made of the GNU Scientific Library (GALASSI *et al.*, 2013) and the GMP (GRANLUND and THE GMP DEVELOPMENT TEAM, 2012) and MPFR (FOUSSE *et al.*, 2007) multiple precision libraries, applying Romberg Integration (BAUER, 1961), is available upon request.

Figure 2: Estimate of power as a function of α when exponential growth is the null hypothesis, and the test statistic is $\sup\{\ell(\vartheta, \xi^{(n)}), \vartheta \in \Theta_0\} - \sup\{\ell(\vartheta, \xi^{(n)}), \vartheta \in \Theta_1\}$ (5), with $\ell(\vartheta, \xi^{(n)})$ the log of the Poisson likelihood function (9) (no lumping) with $n = 200$ and segregating sites (s) as shown. The symbols denote the size of the test as shown in the legend. The interval hypotheses are $\Theta_\beta \equiv \{\beta : \beta \in \{0, 10, \ldots, 1000\}\}$ and $\Theta_\alpha \equiv \{\alpha : \alpha \in \{1, 1.05, \ldots, 2\}\}$.

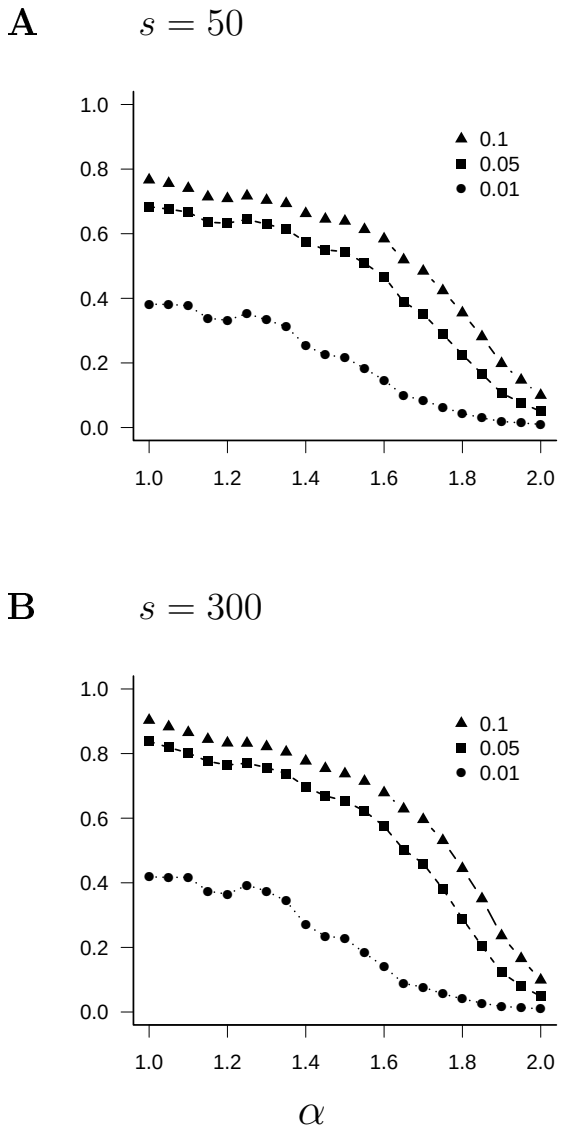


Figure 3: Estimate of power as a function of α when exponential growth is the null hypothesis, and the test statistic is $\sup\{\ell(\vartheta, \xi^{(n)}), \vartheta \in \Theta_0\} - \sup\{\ell(\vartheta, \xi^{(n)}), \vartheta \in \Theta_1\}$ (5), with $\ell(\vartheta, \xi^{(n)})$ the log of the Poisson likelihood function (9) (no lumping). The different symbols denote the size of the test as shown in the legend. The interval hypotheses are $\Theta_\beta \equiv \{\beta : \beta \in \{0, 10, \ldots, 1000\}\}$ and $\Theta_\alpha \equiv \{\alpha : \alpha \in \{1, 1.05, \ldots, 2\}\}$.

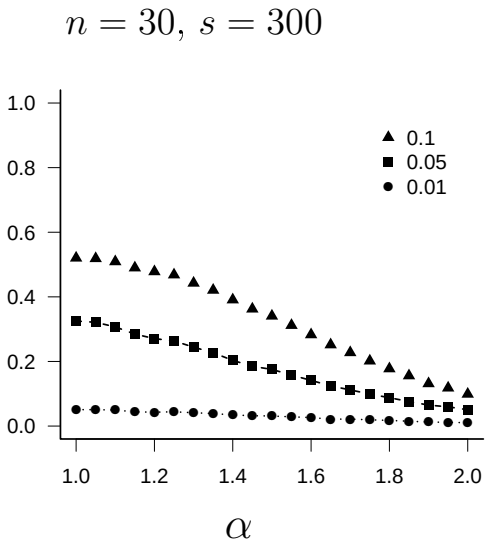
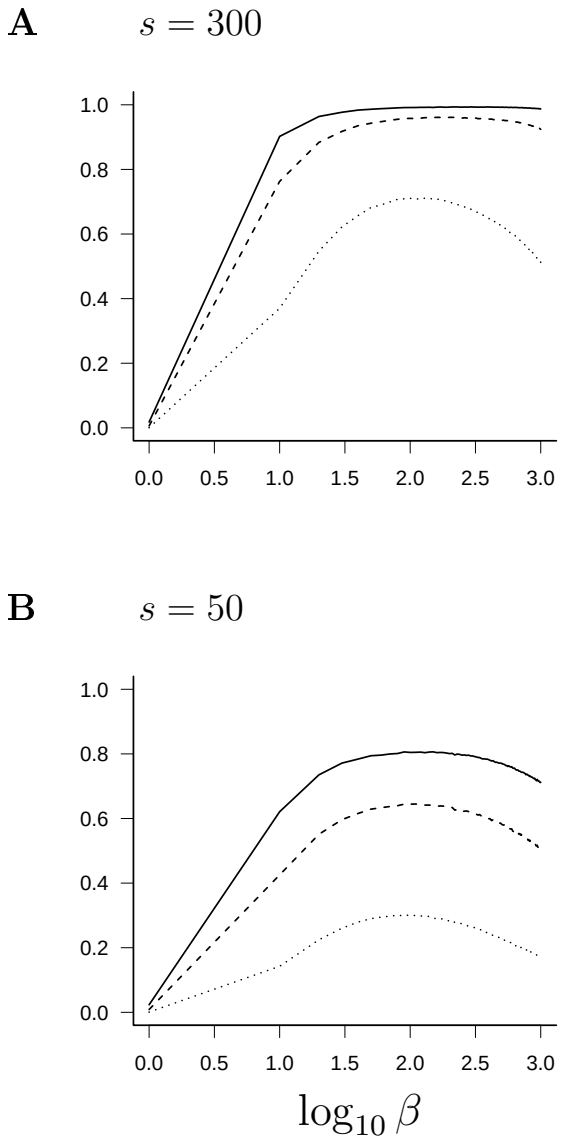


Figure 4: Estimate of power as a function of $\log_{10}\beta$ when the Beta-coalescent is the null hypothesis, and the test statistic is $\sup\{\ell(\vartheta,\xi^{(n)}),\vartheta\in\Theta_0\}-\sup\{\ell(\vartheta,\xi^{(n)}),\vartheta\in\Theta_1\}$ (5), with $\ell(\vartheta,\xi^{(n)})$ the log of the Poisson likelihood function (9) (no lumping); with $n=200$ and number of segregating sites s as shown. The test sizes are 0.1 (solid line), 0.05 (dashed line), 0.01 (dotted line). The interval hypotheses are $\Theta_\beta\equiv\{\beta:\beta\in\{0,10,\ldots,1000\}\}$ and $\Theta_\alpha\equiv\{\alpha:\alpha\in\{1,1.05,\ldots,2\}\}$. Values at $\log_{10}\beta=0$ correspond to the Kingman coalescent.

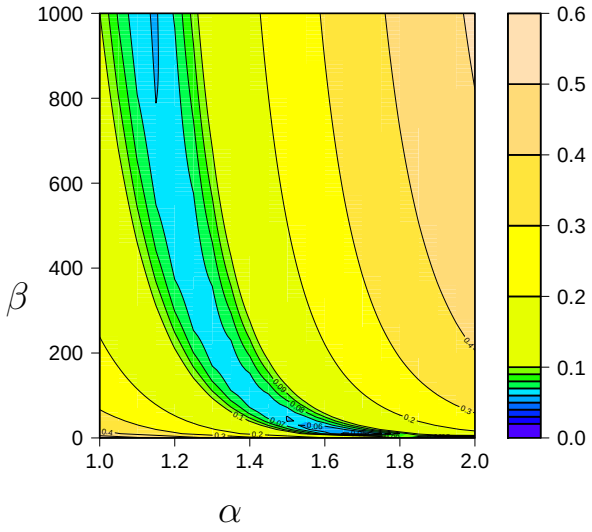


Given additional information about α or β could lead one to test the point hypotheses $H_0 = \{\beta = \beta^*\}$ against $H_1 = \{\alpha = \alpha^*\}$ for some fixed $\beta^* \geq 0$ and $\alpha^* \in [1, 2]$, or vice versa. An example of power estimates for point hypotheses, as a function of sample size n, is shown in Figure S5. High power can be obtained even for relatively small number of segregating sites ($s = 10$). Even excluding the singletons (cf. eg. ACHAZ, 2008) yields high power for some parameter values (Figures S6 and S7).

Considerations of statistical power for point hypotheses naturally lead to the following question: how does the distance between expected site-frequency spectra behave as a function of relevant coalescent parameters? Figure 5 is an effort to understand the relation between the expected SFS for the two models (`E` and `B`) by graphing the distance between expected normalised spectra $\varphi_i^{(n),\mathtt{E}}$ and $\varphi_i^{(n),\mathtt{B}}$ as a function of α and β. The normalised spectrum is of course invariant with respect to mutation rate θ. Figure 5 shows a pattern one would expect from predictions of relative lengths of external branches, and hence singleton polymorphisms - increasing as β increases or α decreases. Similar conclusions can be reached from Figure S4, which shows the distances for $n = 1000$. In particular, one observes that the minimum of the curves for eg. $\beta = 1000$ in Figure S4 shifts as sample size n increases. The smallest ℓ_2 distance in Figure 5 is ≈ 0.02 - excluding 0 due to comparison of Kingman coalescent with itself ($\beta = 0$ and $\alpha = 2$). Naturally one would also want to compare the quantiles of $\xi_i^{(n)}$, or $\zeta_i^{(n)}$, associated with different processes, since two very different distributions can still have the same mean.

Figures 5 and S4 indicate the presence of a region, essentially a curve in the two-dimensional (α, β) parameter space, along which the lowest ℓ_2 distance is reached.

Figure 5: The ℓ_2-distance $\left(\sum_i \left(\varphi_i^{(n),\mathtt{E}} - \varphi_i^{(n),\mathtt{B}}\right)^2\right)^{1/2}$ of the expected normalized spectra $\varphi_i^{(n),\Pi}$ (2) as a function of α and β for $n = 200$. Expected values were computed exactly. The gridpoints are $\alpha \in \{1, 1.05, \ldots, 2\}$, $\beta \in \{0, 10, 20, \ldots, 1000\}$. The smallest distance for the gridpoints chosen is ≈ 0.02 - excluding 0 due to comparison of Kingman coalescent with itself ($\beta = 0$ and $\alpha = 2$).



Mean misclassification probabilities & posterior probabilities for the ABC approach

We analyse how well an ABC approach using the nSFS resp. the folded nSFS and their lumped variants as summary statistic can distinguish between exponential growth and the Beta$(2-\alpha,\alpha)$-coalescent. The distinction ability of the ABC model comparison is assessed as described in the Methods section. We specify the following parameters

- The Beta$(2-\alpha,\alpha)$-coalescent parameter α has the uniform distribution on $[1,2]$ as prior distribution.
- The growth rate β has the uniform distribution on the set $\{0,10,20,\ldots,\beta_{\max}\}$ of subsequent multiples of 10 as prior distribution.
- We simulate the nSFS $n_{reps}=2\times 10^5$ times for $n=200$ in each model class.

See Tables 1 and S1-S3 for the estimates of posterior probabilities and misclassification probabilities (some with one replication).

Table 1: Approximations of the mean posterior probabilities and misclassification probabilities for the ABC model comparison for tolerance $x=0.01$, assumed number $s=60$ of observed mutations and using either the nSFS or the nfSFS as summary statistics. n_{cv} denotes the number of cross-validations 'lumped' denotes which mutation classes are lumped into one class. The maximal growth rate used in all model comparisons is $\beta_{\max}=1000$.

fold	lump	n_{cv}	$\mathbb{E}^{\mathtt{B}}\left[\pi(\mathtt{E}\vert\underline{\zeta})\right]$	$\mathbb{E}^{\mathtt{E}}\left[\pi(\mathtt{B}\vert\underline{\zeta})\right]$	$\mathbb{E}^{\mathtt{B}}\left[\pi(\varrho^{\mathtt{E}/\mathtt{B}}>1\vert\underline{\zeta})\right]$	$\mathbb{E}^{\mathtt{E}}\left[\pi(\varrho^{\mathtt{E}/\mathtt{B}}<1\vert\underline{\zeta})\right]$
no	10+	24000	0.301	0.246	0.258	0.128
no	50+	12000	0.321	0.291	0.262	0.125
no	100+	1200	0.332	0.293	0.282	0.17
yes	10+	24000	0.319	0.253	0.281	0.125
yes	50+	12000	0.34	0.287	0.284	0.155
yes	no	12000	0.343	0.291	0.287	0.162

Appropriate lumping seems to decrease the error probabilities. For $s=60$ observed mutations, strong lumping is decreasing the error probabilities for most parameter choices,

whereas for $s = 300$ (Table S1) moderate lumping seems to decrease them the most. Not surprisingly growth rates closer to zero are harder to distinguish from the Beta$(2-\alpha, \alpha)$-coalescent models than higher growth rates (see Tables S1 and S3). Additionally, a lower count of observed mutations leads to higher error probabilities, as does using the folded nSFS as summary statistics.

Discussion

Distinguishing between multiple merger coalescents and population growth is an important task, in particular since patterns of genetic variation produced by the two demographic effects, and summarized in the site-frequency spectrum, is expected to be similar.

The unit of time of different coalescent processes can vary considerably. As discussed above, predictions about genetic variation can be compared in at least three different ways. One is with time computed in coalescent units, and employing the scaled mutation rate θ. Another way is comparing normalised spectra $\underline{\zeta}^{(n)}$, whose comparison should be independent of timescaling. The third option is to rescale time in generations, which is of course only feasible when one can compute the coalescence probability c_N. Once again, in an ideal haploid Wright-Fisher population, $c_N = 1/N$. In an ideal Schweinsberg population, c_N is proportional to $N^{1-\alpha}$, $1 < \alpha < 2$ (Schweinsberg, 2003). The difference between $1/N$ and $N^{1-\alpha}$ can clearly be substantial. Consequently, it becomes quite hard to compare estimates of θ between different processes, since $\theta/2 \approx \mu/c_N$ for any coalescent process (where μ is the per-generation mutation rate). Thus, our recommendation would be to compare scaled spectra $\underline{\zeta}^{(n)}$.

A key result from our power estimates is that, even for moderate sample size, and based on interval hypotheses, the two processes can be distinguished for significant parts of the parameter space of α and β. By using recent results (Birkner *et al.*, 2013b) on computing $\mathbb{E}^{\Pi}\left[\xi_i^{(n)}\right]$ associated with Lambda-coalescents, and recursions for computing $\mathbb{E}^{\mathbf{E}}\left[\xi_i^{(n)}\right]$ (Supporting Information), allowing us to work within an approximate likelihood framework, we obtain very promising results. Thus, given sample size (n) and observed number of segre-

gating sites (s), our recommendation would be to always estimate power based on interval hypotheses.

The construction of a formal statistical test to distinguish between a multiple merger coalescent process and population growth is complicated in both the minimum-distance and the likelihood framework. A suitable distance-metric would be ($p \geq 1$)

$$Z_p^{(n),\Pi} := \left(\sum_{i=1}^{n-1} \frac{\left| \zeta_i^{(n)} - \mathbb{E}^{\Pi}\left[\zeta_i^{(n)}\right] \right|^p}{\left(\mathrm{Var}^{\Pi}\left[\zeta_{\mathrm{i}}^{(\mathrm{n})}\right] \right)^{p/2}} \right)^{1/p},$$

or a lumped version thereof, where $\mathrm{Var}^{\Pi}\left[\zeta_{\mathrm{i}}^{(\mathrm{n})}\right]$ denotes the variance of $\zeta_i^{(n)}$ associated with coalescent process Π. However, we neither have a closed-form expression for the expected value nor variance of $\zeta_i^{(n)}$ as a function of sample size or coalescence parameters, let alone having any knowledge about the distribution of $Z_p^{(n),\Pi}$. In the likelihood framework, the hypotheses are not nested, and thus it's not clear if convergence to a χ^2-distribution holds. One may instead apply an ABC approach.

Our analysis for sample size $n = 200$ suggests that an ABC method based on the ℓ_2 distance of simulated values to an observed nSFS is able to distinguish between Beta$(2-\alpha, \alpha)$-coalescents or n-coalescents with growth with reasonably low error rates. This holds true at least for a high enough number of observed mutations and models different enough from Kingman's n-coalescent (i.e., β not close to 0, α not close to 2). The nSFS is used because it is more robust to changes in mutation rates than the SFS. We exploit this by estimating the mutation rate from the (assumed) number of observed mutations in the sample instead of drawing it independently from a prior distribution. This is done to reduce the dimensionality of the inference problem.

It is intriguing that using the complete nSFS as summary statistics in the ABC approach yields higher errors than using intermediate resp. strong lumpings of the nSFS. A possible explanation for the positive effect of lumping lies in the relationship between the branch lengths of the coalescent model, the mutation rate and the SFS. Consider the approximate

likelihood function (9). Assume that the distribution of the SFS is approximately composed of independent Poisson distributions with parameter $\frac{\theta}{2}\mathbb{E}^{\Pi}\left[B_i^{(n)}\right]$ for $i \in [n-1]$. For a Poisson-distributed random variable X with parameter κ, we have $\frac{\sqrt{\mathrm{Var}(X)}}{\mathbb{E}(X)} = \frac{1}{\sqrt{\kappa}}$, thus showing that smaller Poisson parameters yield a higher amount of variation relative to their expected value. Thus, classes in the SFS with small underlying branch lengths (which tend to be in the right tail of the SFS) and/or a low mutation rate show relatively more variation compared to their contribution to the total number of mutations than those with longer branches or if the mutation rate is higher. Lumping such classes together, under (9), yields again a Poisson-distributed lumped class, but with Poisson parameter being the sum of parameters from the classes lumped together. Thus, the variation within this class relative to its contribution to the total number of mutations is reduced by lumping. If different coalescent models show different mean behaviour of (lumped) classes, lumping reduces noise and thus increases the chance to correctly identify the underlying model. Naturally, this effect is weakened by higher mutation rates and/or higher sample size n (e.g., consider the limit results for the SFS in Berestycki *et al.* (2013) and Kersting and Stanciu (2013)).

However, the inequalities $\mathbb{E}^A\left[\xi_i^{(n)}\right] > \mathbb{E}^B\left[\xi_i^{(n)}\right]$, and $\mathbb{E}^A\left[\xi_j^{(n)}\right] < \mathbb{E}^B\left[\xi_j^{(n)}\right]$ for $i \neq j$ and two different coalescent models A and B, would, if extreme lumping is applied, eg. collecting together in one group all mutations other than singletons, reduce the signal of the different models. Extreme lumping would thus decrease the chance to correctly identify the model. This heuristic could explain the pattern within the error probabilities. To check whether this heuristic actually holds true, one would need to have better knowledge of the distribution of the SFS resp. nSFS.

Thus, using an appropriate weighing of the variables in the nSFS resp. SFS should improve the power to distinguish between model classes. It would also be a worthwile future study to see whether a one-dimensional summary of the SFS similar to Tajima's D or Fay and Wu's H, as described in Achaz (2009), could yield similar or even higher power to distinguish between the model classes than the complete (possibly reweighted) nSFS. In our ABC computations a simple rejection-based approach was applied. More sophisticated

techniques are available (see BEAUMONT (2010) for an overview) that may improve the prediction accuracy.

Marine organisms with external fertilization and Type III survivorship curves, such as Pacific oysters or Atlantic cod, and in which each gender produces a large number of gametes (MAY, 1967; STRATHMANN, 1987), are prime candidates for natural populations exhibiting high fecundity and skewed offspring distributions (BECKENBACH, 1994; BOOM *et al.*, 1994; ÁRNASON, 2004). Previous analysis of mtDNA of Atlantic cod (cf. eg. BIRKNER *et al.*, 2013b) and Pacific oysters (SARGSYAN and WAKELEY, 2008) supports the hypothesis that multiple merger coalescents may be more appropriate than the Kingman coalescent as baseline models for high fecundity organisms with skewed offspring distributions. Our ABC analysis of the Atlantic cod data of ÁRNASON (2004) (see Supporting Information) indicates exponential growth is slightly favored over the Beta$(2-\alpha, \alpha)$-coalescent. ÁRNASON (2004) does exclude population growth in his analysis. Our results indicate that the SFS information one obtains from the ÁRNASON (2004) data may not have enough polymorphic sites to distinguish between exponential population growth and the Beta$(2-\alpha, \alpha)$-coalescent.

The problem of distinguishing between multiple merger coalescents and population growth was our motivation. We proceeded by analysing two natural models with clear biological interpretation, the Beta$(2-\alpha, \alpha)$-coalescent and exponential growth using the site-frequency spectrum. The power estimation results are very promising: for even moderate sample size we can distinguish between the two models for large parts of the associated parameter spaces with high power. Thus, we recommend, for a given sample size (n) and number of segregating sites (s) to estimate power based on interval hypotheses. Approximate Bayesian computation also yields convincing results, at least for a high level of polymorphism. The SFS, in general, has enough information to distinguish between exponential growth and multiple merger coalescent processes. Comparing our single-locus methods and results to those obtained for multiple loci remains an important future exercise.

Acknowledgements: J. Blath and B. Eldon were supported by Deutsche Forschungsgemeinschaft (DFG) grant BL 1105/3-1, and M. Birkner by DFG grant BI 1058/2-1, as parts

of SPP Priority Programme 1590. F. Freund thanks Luca Ferretti and Guillaume Achaz (SMILE, Collège de France, Paris) for discussions about the site-frequency spectrum.

References

Abramowitz, M., and I. A. Stegun, editors, 1964 *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Number 55 in Applied Mathematics Series. National Bureau of Standards, Washington, D.C.

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.

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

Baragatti, M., and P. Pudlo, 2014 An overview on approximate bayesian computation. ESAIM **44**: 291–299.

Bauer, F. L., 1961 Algorithm 60 – Romberg Integration. Communications ACM **4**: 255.

Beaumont, M. A., 2010 Approximate bayesian computation in evolution and ecology. Annual Review of Ecology, Evolution, and Systematics **41**: 379–406.

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

Berestycki, J., N. Berestycki, and V. Limic, 2013 A sampling formulae for Lambda-coalescents. To appear (arXiv:1201.6512).

Berkson, J., 1980 Minimum chi-square, not maximum likelihood! Annals of Statistics **8**: 457–487.

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

Birkner, M., J. Blath, and B. Eldon, 2013b Statistical properties of the site-frequency spectrum associated with lambda-coalescents. Genetics **195**: 1037–1053.

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.

Boom, J. D. G., E. G. Boulding, and A. T. Beckenbach, 1994 Mitochrondrial DNA variation in introduced populations of Pacific Oyster, *Crassostrea gigas*, in British Columbia. Can J Fish Aquat Sci **51**: 1608–1614.

Cannings, C., 1974 The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. Adv Appl Prob **6**: 260–290.

Cannings, C., 1975 The latent roots of certain Markov chains arising in genetics: a new approach. II. Further haploid models. Adv Appl Prob **7**: 264–282.

Carr, S. M., and H. D. Marshall, 2008 Intraspecific phylogeographic genomics from multiple complete mtDNA genomics in Atlantic cod (*Gadus morhua*): origins of 'codmother', transatlantic vicariance, and midglacial population expansion. Genetics **180**: 381–389.

Chen, W.-C., 2011 *Overlapping Codon Model, Phylogenetic Clustering, and Alternative Partial Expectation Conditional Maximization Algorithm*. Ph.D. thesis, Iowa State University, Ames, Iowa. Http://gradworks.umi.com/34/73/3473002.html.

Csilléry, K., O. François, and M. G. B. Blum, 2012 ABC: an R package for approximate bayesian computation (ABC). Methods in Ecology and Evolution **3**: 475–479.

Cucala, L., and J. Marin, 2013 Bayesian inference on a mixture model with spatial dependence. J Comp Graph Stats **22**: 584–597.

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

Donnelly, P., and S. Tavaré, 1995 Coalescents and genealogical structure under neutrality. Ann Rev Genet **29**: 401–421.

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.

Enge, A., M. Gastineau, P. Théveny, and P. Zimmermann, 2012 *mpc — A library for multiprecision complex arithmetic with exact rounding*. INRIA, 1.0 edition. `http://mpc.multiprecision.org/`.

Fay, J. C., and C.-I. Wu, 2000 Hitchhiking under positive darwinian selection. Genetics **155**: 1405–1413.

Fousse, L., G. Hanrot, V. Lefèvre, P. Pélissier, and P. Zimmermann, 2007 MPFR: A multiple-precision binary floating-point library with correct rounding. ACM Transactions on Mathematical Software **33**: 13:1–13:15.

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

Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133**: 693–709.

Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, *et al.*, 2013 *GNU Scientific Library Reference Manual*, third edition. ISBN 0954612078.

Granlund, T., and the GMP development team, 2012 *GNU MP: The GNU Multiple Precision Arithmetic Library*, 5.0.5 edition.

Griffiths, R. C., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Comm Statistic 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.

Hein, J., M. H. Schierup, and C. Wiuf, 2005 *Gene genealogies, variation and evolution.* Oxford University Press, Oxford, UK.

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

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

Hudson, R. R., 1990 Gene genealogies and the coalescent process. In D. J. Futuyma and J. Antonovics, editors, *Oxford surveys in evolutionary biology*, volume 7. Oxford University Press, Oxford, 1–44.

Hudson, R. R., 2002 Generating samples under a wright-fisher neutral model. Bioinformatics **18**: 337–338.

Jeffreys, H., 1961 *Theory of Probability*. Oxford University Press, Oxford, UK, 3rd edition.

KAJ, I., and S. KRONE, 2003 The coalescent process in a population with stochastically varying size. J Appl Probab **40**: 33–48.

KASS, R. E., and A. E. RAFTERY, 1995 Bayes factors. Journal of the American Statistical Association **90**: 773–795.

KERNIGHAN, B. W., and D. M. RITCHIE, 1988 *The C programming language*. Prentice Hall, Englewood Cliffs, New Jersey, second edition.

KERSTING, G., and I. STANCIU, 2013 The internal branch lengths of the Kingman coalescent. To appear (arXiv:1303.4562).

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

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

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

KOSKELA, J., P. JENKINS, and D. SPANÒ, 2013 Computational inference beyond Kingman's coalescent. submitted .

MAY, A. W., 1967 Fecundity of Atlantic cod. J Fish Res Brd Can **24**: 1531–1551.

MILLAR, P. W., 1984 A general approach to the optimality of minimum distance estimators. Trans. Amer. Math. Soc. **286**: 377–418.

MÖHLE, M., 1998 Robustness results for the coalescent. J Appl Probab **35**: 438–447.

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

MYERS, S., C. FEFFERMAN, and N. PATTERSON, 2008 Can one learn history from the allelic spectrum? Theor Popul Biol **73**: 342–348.

NORDBORG, M., 2001 Coalescent theory. In D. J. Balding, M. J. Bishop and C. Cannings, editors, *Handbook of statistical genetics*, chapter 25. John Wiley & Sons, Chichester, UK, 2nd edition, 179–212.

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

POLANSKI, A., A. BOBROWSKI, and M. KIMMEL, 2003 A note on distribution of times to coalescence, under time-dependent population size. Theor Popul Biol **63**: 33–40.

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. Mol Biol Evol **16**: 1791–1798.

R CORE TEAM, 2012 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

RAMÍREZ-SORIANO, A., S. E. RAMOS-ONSINS, J. ROZAS, F. CALAFELL, and A. NAVARRO, 2008 Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics **179**: 555–567.

RAMOS-ONSINS, S. E., and J. ROZAS, 2002 Statistical properties of new neutrality tests against population growth. Mol Biol Evol **19**: 2092–2100.

ROGERS, A. R., and H. C. HARPENDING, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol **9**: 552–569.

RUBIN, D. B., 1984 Bayesian justifiable and relevant frequency calculations for the applied statistician. Ann Stats **12**: 1151–1172.

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

SANO, A., and H. TACHIDA, 2005 Gene genealogy and properties of test statistics of neutrality under population growth. Genetics **169**: 1687–1697.

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 Prob **5**: 1–50.

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

SLATKIN, M., and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129**: 555–562.

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.

STOEHR, J., P. PUDLO, and L. CUCALA, 2014 Geometric summary statistics for ABC model choice between hidden Gibbs random fields. arXiv:1402.1380 .

STRATHMANN, M. F., 1987 *Reproduction and development of marine invertebrates of the northern Pacific coast*. U. Washington Press.

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

TAJIMA, F., 1989a The effect of change in population size on DNA polymorphism. Genetics **123**: 597–601.

TAJIMA, F., 1989b Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.

TAVARÉ, S., D. J. BALDING, R. C. GRIFFITHS, and P. DONNELLY, 1997 Inferring coalescence times from DNA sequence data. Genetics **145**: 505–518.

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

WAKELEY, J., 2007 *Coalescent theory*. Roberts & Co, Greenwood Village.

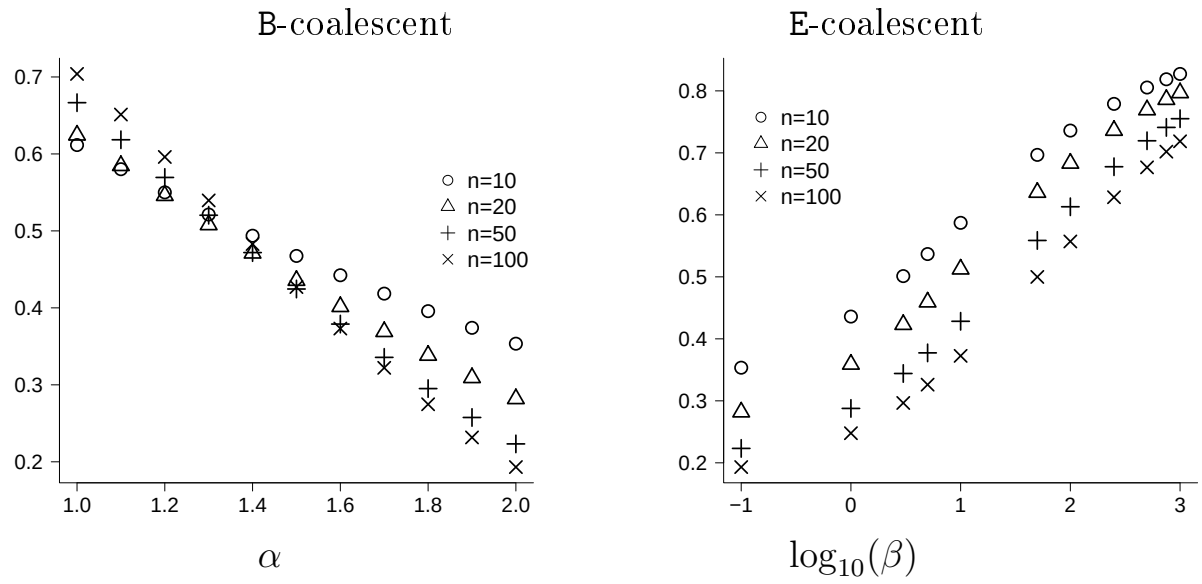
WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor Pop Biol **7**: 1539–1546.

Supporting Information

Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents?

M. Birkner, J. Blath, B. Eldon, F. Fabian

Figure S1: Relative expected length of external branches $\varphi_1^{\Pi}(n)$ (2) for $\Pi \in \{\mathtt{B}, \mathtt{E}\}$ as a function of the coalescent parameters α resp. β; '2.0' and '-1' denote the Kingman coalescent; number of leaves n as shown. Expected values were computed exactly.



The SFS for multiple merger coalescents

Some basic theory of multiple merger coalescents, which are also known as *Lambda*-coalescents in the mathematics literature will now be briefly reviewed. The theory associated with *simultaneous multiple mergers*, or so-called Ξ-coalescents (Schweinsberg, 2000; Möhle and Sagitov, 2001), could be treated by similar methods, but will not be discussed at this point.

The discussion of the relevance of Lambda-coalescents in population genetics started with Eldon and Wakeley (2006). Recall that a Lambda-coalescent, formally introduced by Pitman (1999); Sagitov (1999), and Donnelly and Kurtz (1999), is a partition-valued exchangeable coalescent process determined by a finite measure Λ on $[0,1]$. When Λ is associated with the beta-distribution with parameters $2-\alpha$ and α for $1 \leq \alpha < 2$ (Schweinsberg, 2003), any particular subset of $k \in \{2,\ldots,b\}$ blocks merges into one (ie. a particular set of k labelled ancestral lineages coalesce) at rate $\lambda_{b,k} = B(k-\alpha, n-k+\alpha)/B(2-\alpha,\alpha)$, where $B(\cdot,\cdot)$ is the beta function. In contrast, $\lambda_{b,k} = \mathbf{1}_{(k=2)}$ for the Kingman coalescent ($\Lambda(dx) = \delta_0(dx)$). The Beta$(2-\alpha,\alpha)$-coalescent introduces a *coalescent* parameter α, which can be estimated from genetic data (Eldon, 2011; Birkner *et al.*, 2013b; Birkner and Blath, 2008; Steinrücken *et al.*, 2013).

As before, for any Lambda-coalescent $\Pi^{(\Lambda)}$ and $i \in [n-1]$, the expected frequency spectrum $\mathbb{E}^{\Pi^{(\Lambda)}}[\xi_i^{(n)}]$ is given by (1). Denote by $\left(Y_t^{(n),\Pi}, t \geq 0\right)$ the block-counting process, ie. $Y_t^{(n),\Pi}$ gives the number of ancestral lineages active at time t; with $\mathbb{P}\left(Y_0^{(n),\Pi} = n\right) = 1$. In contrast to the Kingman case, $Y^{(n),\Pi^{(\Lambda)}}$ might not hit all possible states between n and 2 when associated with a multiple merger coalescent. An additional implicit conditioning must therefore be employed. Indeed, the term $p^{(n),\Pi}[k,i] = p^{(n),\Pi^{(\Lambda)}}[k,i]$ in the corresponding version of (1) denotes the probability that, in a Lambda-coalescent started from n unlabelled lineages, *conditional* on hitting a state with k lineages, a given one of the k blocks subtends exactly $i \in [n-1]$ leaves.

While the recursive approach underlying (1) can be generalized to compute variances and covariances (Fu, 1995; Birkner *et al.*, 2013b), and in fact to any mixed moment of any order (however, at the cost of rapidly increasing complexity), the explicit distribution of the SFS is

in general unknown. BERESTYCKI *et al.* (2013) provide the large n almost sure asymptotic behaviour of the SFS for Beta-coalescents (and, more generally, all coalescents with a suitable polynomial singularity at 0). In KERSTING and STANCIU (2013), the convergence of the joint law of the internal branch lengths, as n grow large, to a multivariate Gaussian distribution is established in the case of the Kingman coalescent, but the convergence appears to be rather slow. The corresponding result for (subclasses of) Lambda-coalescents is desirable, but currently still elusive. Again the rate of convergence into the scaling limit is expected to be rather slow, and the result might therefore be more of theoretical value.

The SFS under variable population size

The effect of fluctuations in population size on the underlying ancestry has been investigated in various articles, see in particular GRIFFITHS and TAVARÉ (1998), who derive an analog of (1), and KAJ and KRONE (2003) who link the Wright-Fisher approximation (with fluctuating population size) with the limiting genealogy. An important point is that (moderate) exogenously determined fluctuations in population size do not affect the binary coalescence structure of the genealogy, but enter only in the distribution of the relative length of the internal branches, and hence the genealogy can be described by a time-change of a Kingman coalescent. In contrast, *large* fluctuations, in particular very severe bottlenecks, can lead to very different coalescent models, see (BIRKNER *et al.*, 2009).

Recursions for the expected values and covariances of the site-frequency spectrum associated with moderate fluctuations in population size will now be obtained. The recursion can be obtained following POLANSKI *et al.* (2003). We consider a time-inhomogeneous Kingman coalescent, started in n lineages, where each pair of lines present at time $t \geq 0$ merges at a rate $\nu(t)$ (instead of rate 1) (cf. GRIFFITHS and TAVARÉ, 1998). Given the exogeneously determined time-change $\nu = (\nu(t), t \geq 0)$, the expected frequency spectrum $\mathbb{E}[\xi_i^{(n),\nu}], i \in [n-1]$, is again of the form (1), and the time-change ν enters only in the distribution of the $T_k^{(n)} = T_k^{(n),\nu}$, $2 \leq k \leq n$, that is, the distribution of the lengths of the time intervals of the block-counting process $Y_t^{(n),\nu}$ during which there are exactly k lineages.

To evaluate $\mathbb{E}[\xi_i^{(n),\nu}]$ one needs information about $\mathbb{E}[T_k^{(n),\nu}]$. Define

$$S_j^{(n),\nu} := T_n^{(n),\nu} + T_{n-1}^{(n),\nu} + \cdots + T_j^{(n),\nu}, \quad j = n, \ldots, 2 \tag{S1}$$

to be the time at which the block counting process $Y^{(n),\nu}$ jumps from j to $j-1$ lineages (with the convention $S_{n+1}^{(n),\nu} := 0$). Further, abbreviate, for $t \geq 0$ and $j \in 2, \ldots, n$,

$$F(t) := \int_0^t \nu(u)\, du \quad \text{and} \quad a_j := \int_0^\infty e^{-\binom{j}{2} F(s)}\, ds, \tag{S2}$$

assuming that the first integral in (S2) is finite. It is possible to compute the marginal density of $S_m^{(n),\nu}$ using the well-known fact that the density of a convolution of exponentials with different rates can be written as a linear combination of exponential densities (Polanski *et al.*, 2003). Indeed, this leads to the following recursive characterization for the expected values of the $S_k^{(n),\nu}$ and $T_k^{(n),\nu}$ (see section 'Deriving the expected SFS under variable population size' below for a proof): For $2 \leq m \leq n$ and $m \leq j \leq n$ let

$$c_m^{(j,n)} := \prod_{\substack{m \leq i \leq n \\ i \neq j}} \frac{\binom{i}{2}}{\binom{i}{2} - \binom{j}{2}} = (-1)^{j-m} \frac{(2j-1)m}{j(j-1)} \frac{\binom{n}{j}\binom{j+m-2}{j}\binom{j}{m}}{\binom{n+j-1}{j}}, \tag{S3}$$

(put $c_m^{(j,n)} = 0$ for $j < m$). Then we have the representation

$$\mathbb{E}\big[S_m^{(n),\nu}\big] = \sum_{j=m}^{n} c_m^{(j,n)} a_j\,, \tag{S4}$$

and

$$\mathbb{E}\big[T_k^{(n),\nu}\big] = \mathbb{E}\big[S_k^{(n),\nu} - S_{k+1}^{(n),\nu}\big] = c_k^{(k,n)} a_k + \sum_{j=k+1}^{n} \big(c_k^{(j,n)} - c_{k+1}^{(j,n)}\big) a_j\,, \tag{S5}$$

which allows us to compute the expectation of $\xi_i^{(n)}$ easily from (1):

$$\mathbb{E}\big[\xi_i^{(n),\nu}\big] = \frac{\theta}{2}\sum_{k=2}^{n} k \cdot p_{n,k}(i) \cdot \left[c_k^{(k)} a_k + \sum_{j=k+1}^{n} \big(c_k^{(j)} - c_{k+1}^{(j)}\big) a_j\right]. \tag{S6}$$

Again we emphasize that the $p_{n,k}(i)$ are not affected by population growth, and can be given explicitly (exactly as in the Kingman-case, cf. Fu (1995)):

$$p_{n,k}(i) := \mathbf{1}_{\{k \le n-i+1\}} \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}}.$$

We now specify the main ingredient a_j (depending on $F(t), t \ge 0$ and hence $\nu(t), t \ge 0$) explicitly for two important special cases:

a) Exponential growth. In the case of an exponentially growing population with growth parameter β, that is, $\nu(t) = e^{\beta t}$,

$$a_j = \frac{1}{\beta} \exp\Big(\beta^{-1}\tbinom{j}{2}\Big) E_1\Big(\beta^{-1}\tbinom{j}{2}\Big), \tag{S7}$$

where

$$E_1(t) := \int_t^\infty \frac{e^{-x}}{x}\, dx = \int_1^\infty \frac{e^{-tx}}{x}\, dx \tag{S8}$$

is an exponential integral function, c.f. e.g. (Abramowitz and Stegun, 1964, 5.1.1).

Two practical issues must be stated at this point. The coefficients $c_m^{(j,n)}$ (S3) involve binomial coefficients which become large for large sample sizes (number of leaves), which can lead to numerical errors. Multiple precision packages such as GNU MPC (Enge *et al.*, 2012) or MPFR (Fousse *et al.*, 2007) can be employed to compute the $c_m^{(j,n)}$ coefficients exactly. In addition, one must evaluate numerically the integral $E_1(t)$ in (S8). The computing time of $E_1(t)$ using multiple precision packages quickly increases with sample size. Hence, for large sample sizes ($n > 200$), we applied simulations to estimate expected values associated with exponential population growth using the results of Griffiths and Tavaré (1998), specifically Equation (2.6), which gives a simple way of drawing values of the $S_j^{(n),\nu}$, the

successive coalescence times.

b) Algebraic ('power law') growth. In the case of algebraic growth of the form $\nu(t) = t^\gamma$ for some $\gamma > 0$,

$$a_j = \frac{\Gamma\big(1/(\gamma+1)\big)}{\gamma^{\gamma/(\gamma+1)}} \binom{j}{2}^{-1/(\gamma+1)}. \tag{S9}$$

While exponential growth is a natural model for a population described by a supercritical branching mechanism, this appears less natural in the power-law case.

Based on Equation (23) in Fu (1995), it is also possible to compute the *variance* and the *covariances* of the SFS based on expressions for $\mathbb{E}_\nu[T_k^{(n),\nu} T_l^{(n),\nu}]$, $2 \le k, l \le n$, which in turn can be obtained from

$$\mathbb{E}_\nu[T_k^{(n),\nu} T_l^{(n),\nu}] = \mathbb{E}_\nu[S_k^{(n),\nu} S_l^{(n),\nu}] - \mathbb{E}_\nu[S_{k-1}^{(n),\nu} S_l^{(n),\nu}] - \mathbb{E}_\nu[S_k^{(n),\nu} S_{l-1}^{(n),\nu}] + \mathbb{E}_\nu[S_{k-1}^{(n),\nu} S_{l-1}^{(n),\nu}],$$

noting that, in the above notation,

$$\mathbb{E}\big[(S_m^{(n),\nu})^2\big] = \int_0^\infty s_m^2 \sum_{j=m}^{n} c_m^{(j,n)} \nu(s_m) \binom{j}{2} e^{-\binom{j}{2} F(s_m)}\, ds_m,$$

and

$$\mathbb{E}[S_m^{(n),\nu} S_k^{(n),\nu}] = \mathbb{E}\big[\mathbb{E}[S_m^{(n),\nu} | S_k^{(n),\nu}] S_k^{(n),\nu}\big],$$

where $\mathbb{E}[S_m^{(n),\nu} | S_k^{(n),\nu} = s_k]$ can be computed (it is the expectation under a regular conditional probability) as in (S4) replacing ν by $\tilde{\nu}(\cdot) := \nu(\cdot + s_k)$, $c_m^{(j,n)}$ by $\tilde{c}_m^{(j)} := c_m^{(j,k)}$ and F by $\tilde{F}(\cdot) = F(s_k + \cdot) - F(s_k)$.

Formula (S2) for the (a_j) invites a question analogous to Myers *et al.* (2008), namely whether different choices of F can lead to the same sequence of coefficients (a_j), which is however outside the scope of the present work. Besides such purely theoretical consider-

ations, a rather large body of work has also been devoted to practical inference of population growth from genetic data (see e.g. TAJIMA (1989a), SLATKIN and HUDSON (1991), ROGERS and HARPENDING (1992), and RAMOS-ONSINS and ROZAS (2002)). Simulation-based work include RAMÍREZ-SORIANO *et al.* (2008), who consider the statistical power of several tests under population size increase and decrease, and the impact of recombination. RAMOS-ONSINS and ROZAS (2002) consider the statistical power of statistics based on the site-frequency spectrum to distinguish deterministic population growth from the Kingman coalescent.

Comparison of $\varphi_i^{(n),\mathtt{E}}$ and $\mathbb{E}^{\mathtt{E}}\left[Z_i^{(n)}\right]$

The agreement between $\varphi_i^{(n),\mathbb{E}}$ (2) and $\mathbb{E}^{\mathtt{E}}\left[Z_i^{(n)}\right]$, where $Z_i^{(n)} \equiv B_i^{(n)}/B^{(n)}$ are the relative branch lengths subtending $i \in [n-1]$ leaves (equivalent to DNA sequences), with $B^{(n)} \equiv B_1^{(n)} + \cdots + B_{n-1}^{(n)}$ denoting the total tree length, is checked in Figure S2. In Figure S2, the difference $\left(\varphi_i^{(n),\mathtt{E}} - \overline{Z}_i^{(n)}\right)$ relative to $\varphi_i^{(n),\mathtt{E}}$ is graphed for each class (i), where the $\overline{Z}_i^{(n)}$ were estimated by simulation. Indeed, the agreement between $\varphi_i^{(n),\mathtt{E}}$ and the estimated values $\overline{Z}_i^{(n)}$ is quite good for the large range of β considered, and improves as n increases. The results clearly indicate that agreement between $\mathbb{E}^{\mathtt{E}}[\zeta_i^{(n)}]$ and $\varphi_i^{(n),\mathbb{E}}$ will also be good.

Figure S2: Comparing $\varphi_i^{(n),\mathbf{E}}$ and simulated values $\overline{Z}_i^{(n)} \equiv \overline{B_i^{(n)}/B^{(n)}}$ for exponential growth by graphing $\left(\varphi_i^{(n),\mathbf{E}} - \overline{Z}_i^{(n)}\right)/\varphi_i^{(n),\mathbf{E}}$ against i for n and the exponential growth parameter β as shown, where $B_i^{(n)}$ is total length of branches subtending $i \in [n-1]$ leaves, and $B^{(n)}$ is the total length. Values for $\overline{Z}_i^{(n)}$ were obtained from 10^6 replicates, and $\varphi_i^{(n),\mathbf{E}}$ was computed exactly.

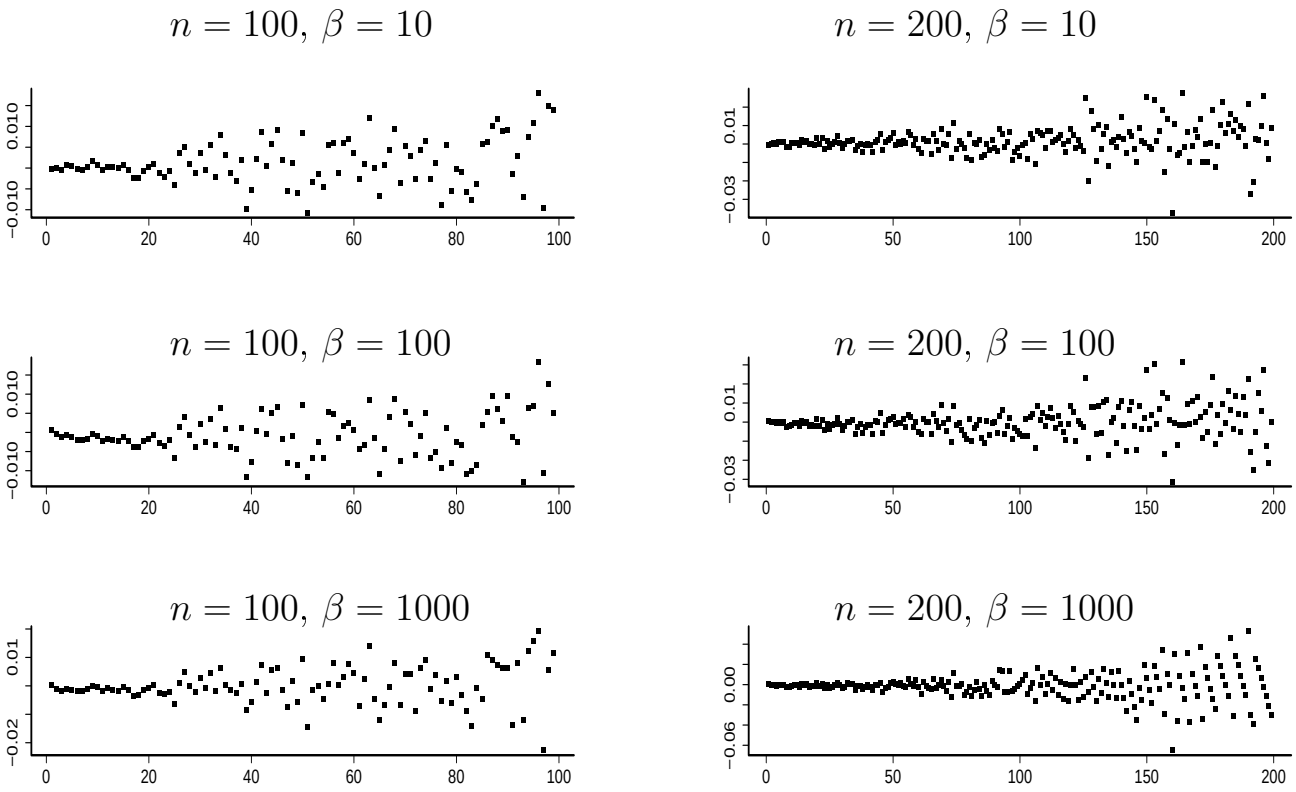
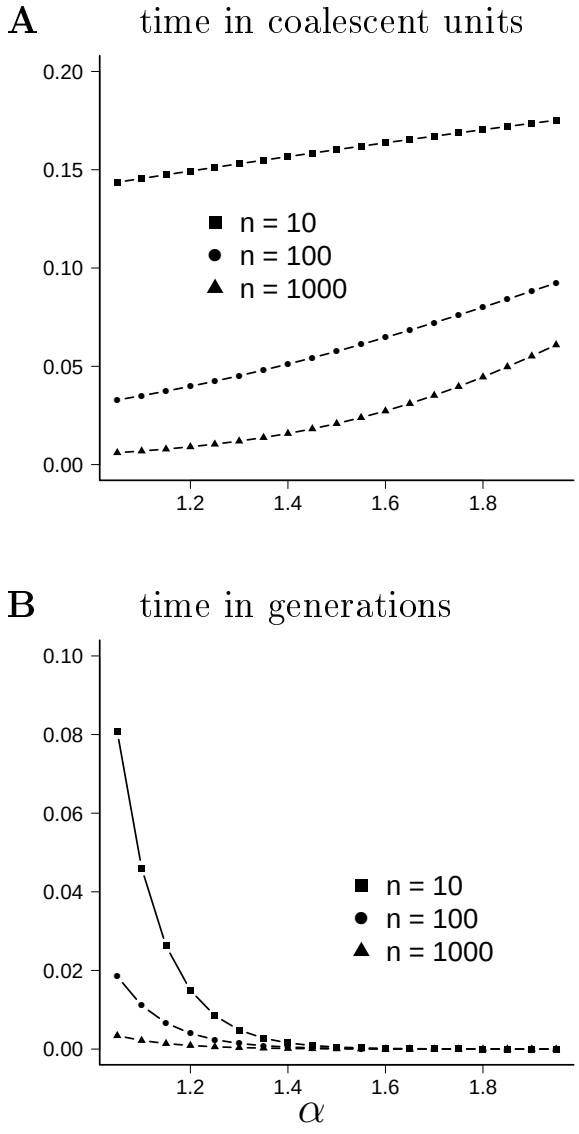


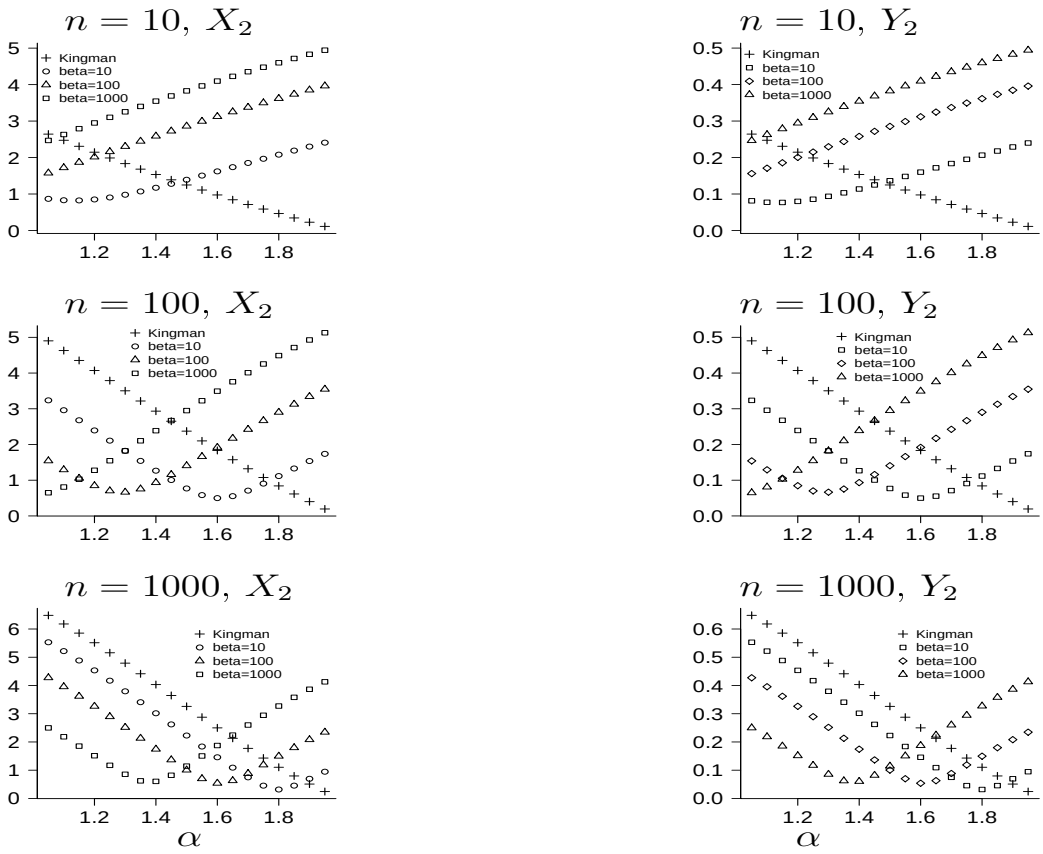
Figure S3: Graphs of $1/\mathbb{E}^{\mathsf{B}}\left[B^{(n)}\right]$, the estimated value of $\theta/2$ per observed mutation when using the Watterson estimator (3) as a function of α (**A**), compare with (3); and the estimated value of μ per observed mutation (**B**), using (4) together with (3), and assuming the timescale $c_N = N^{1-\alpha}$. The number of leaves n are as shown. In **B**, time is converted into generations by multiplying $\mathbb{E}^{\mathsf{B}}\left[B^{(n)}\right]$ with $N^{\alpha-1}$, when $N = 10^5$. One obtains $1/\mathbb{E}^{\mathsf{K}}\left[B^{(n)}\right] = 0.177$, 0.097, and 0.067 coalescent units for $n = 10$, 100, and 1000, resp.



Comparison of $\mathbb{E}^{\mathtt{E}}[\xi_i^{(n)}]$ and $\mathbb{E}^{\mathtt{B}}[\xi_i^{(n)}]$

Exact computations of expected branch lengths $\mathbb{E}^{\mathtt{E}}[B_i^{(n)}]$ associated with exponential growth become inefficient as sample size increases, due to numerical computation of (S8). One can, however, estimate $\mathbb{E}^{\mathtt{E}}[B_i^{(n)}]$ using simulations, applying results from Griffiths and Tavaré (1998). Figure S4 shows the distance between $\mathbb{E}^{\mathtt{E}}[\xi_i^{(n)}]$ and $\mathbb{E}^{\mathtt{B}}[\xi_i^{(n)}]$ as measued by the X_2 norm (S12), and the distance between the normalized spectrum $\varphi_i^{(n),\mathtt{E}}$ and $\varphi_i^{(n),\mathtt{B}}$ as measured by the Y_2 norm (S13). Figure S4 corresponds to certain 'one-dimensional slices' through a two-dimensional graph that would be the analogue of Figure 5 for $n = 1000$.

Figure S4: Graphs of X_2 (S12) and Y_2 (S13) as a function of α when the 'data' is $\mathbb{E}^{\mathtt{E}}\left[\xi_i^{(n)}\right]$ resp. $\varphi_i^{(n),\mathtt{E}}$ (+ for $\mathbb{E}^{\mathtt{K}}$ resp. $\varphi_i^{(n),\mathtt{K}}$) compared to $\mathbb{E}^{\mathtt{B}}\left[\xi_i^{(n)}\right]$ resp. $\varphi_i^{(n),\mathtt{G}}$ for β ($=\mathtt{b}$) and sample size n as shown. Estimates for $\mathbb{E}^{\mathtt{E}}$ were obtained from 10^5 replicates.



Mean misclassification probabilities and posterior probabilities for ABC approach - alternative parameter choices

Table S1: Approximations of mean posterior probabilities and misclassification probabilities for the ABC model comparison for different growth parameter ranges or tolerance rates. The nSFS is used as summary statistics. $\beta_{\max}$ denotes the maximal growth rate used in the growth model, n_{cv} denotes the number of cross-validations; 'lump' denotes which mutation classes are lumped into one class. An expected number $s = 300$ of mutations are assumed.

$\beta_{\max}$	lump	n_{cv}	tolerance	$\mathbb{E}^{\mathtt{B}}\left[\pi(\mathtt{E}\|\underline{\zeta})\right]$	$\mathbb{E}^{\mathtt{E}}\left[\pi(\mathtt{B}\|\underline{\zeta})\right]$	$\mathbb{E}^{\mathtt{B}}\left[\pi(\varrho^{\mathtt{E}/\mathtt{B}} > 1\|\underline{\zeta})\right]$	$\mathbb{E}^{\mathtt{E}}\left[\pi(\varrho^{\mathtt{E}/\mathtt{B}} < 1\|\underline{\zeta})\right]$
10^3	10+	24000	0.01	0.236	0.111	0.181	0.04
"	"	"	"	0.24	0.11	0.181	0.039
10^3	50+	12000	0.01	0.225	0.087	0.183	0.026
"	"	"	"	0.227	0.088	0.188	0.028
10^3	100+	1200	0.01	0.217	0.085	0.193	0.028
"	"	12000	"	0.222	0.08	0.2	0.025
10^3	no	12000	0.01	0.301	0.142	0.232	0.042
"	"	"	"	0.302	0.143	0.234	0.041
500	10+	24000	0.01	0.264	0.132	0.2	0.054
500	50+	12000	0.01	0.24	0.101	0.195	0.038
500	100+	1200	0.01	0.257	0.09	0.222	0.028
100	10+	24000	0.01	0.306	0.208	0.235	0.122
100	50+	12000	0.01	0.274	0.176	0.205	0.101
10^3	10+	24000	0.0025	0.2	0.108	0.149	0.045
10^3	50+	12000	0.0025	0.187	0.08	0.146	0.027
10^3	100+	1200	0.0025	0.182	0.081	0.16	0.03
10^3	no	1200	0.0025	0.246	0.134	0.19	0.049

Table S2: Approximations of mean posterior probabilities and misclassification probabilities for the ABC model comparison ($n = 200$) for alternative tolerance $x = 0.0025$ and assumed expected number $s = 60$ of mutations. The nSFS is used as summary statistics. n_{cv} denotes the number of cross-validations 'lumped' denotes which mutation classes are lumped into one class. The maximal growth rate used in all model comparisons is $\beta_{\max} = 1000$.

lump	n_{cv}	$\mathbb{E}^{\mathtt{B}}\left[\pi(\mathtt{E}\vert\underline{\zeta})\right]$	$\mathbb{E}^{\mathtt{E}}\left[\pi(\mathtt{B}\vert\underline{\zeta})\right]$	$\mathbb{E}^{\mathtt{B}}\left[\pi(\varrho^{\mathtt{E}/\mathtt{B}} > 1\vert\underline{\zeta})\right]$	$\mathbb{E}^{\mathtt{E}}\left[\pi(\varrho^{\mathtt{E}/\mathtt{B}} < 1\vert\underline{\zeta})\right]$
10	24000	0.278	0.245	0.228	0.14
50	12000	0.31	0.258	0.252	0.141
100	12000	0.328	0.265	0.277	0.152
no	12000	0.339	0.262	0.287	0.146

Table S3: Approximations of mean posterior probabilities and misclassification probabilities for the ABC model comparison ($n = 200$) for alternative tolerance $x = 0.001$, assumed expected number $s = 60$ of mutations and alternative prior ranges and distributions. The nSFS is used as summary statistics. n_{cv} denotes the number of cross-validations 'lumped' denotes which mutation classes are lumped into one class. For growth rate β, the prior is uniformly distributed on $\{\beta_{\min}, \beta_{\min} + 10, \ldots, \beta_{\max}\}$. For coalescent parameter α, the prior is uniformly distributed on $[\alpha_{\min}, \alpha_{\max}]$

lump	n_{cv}	$\beta_{\min}, \beta_{\max}$	$\alpha_{\min}, \alpha_{\max}$	$\mathbb{E}^{\mathtt{B}}\left[\pi(\mathtt{E}\vert\underline{\zeta})\right]$	$\mathbb{E}^{\mathtt{E}}\left[\pi(\mathtt{B}\vert\underline{\zeta})\right]$	$\mathbb{E}^{\mathtt{B}}\left[\pi(\varrho^{\mathtt{E}/\mathtt{B}} > 1\vert\underline{\zeta})\right]$	$\mathbb{E}^{\mathtt{E}}\left[\pi(\varrho^{\mathtt{E}/\mathtt{B}} < 1\vert\underline{\zeta})\right]$
10	24000	0,100	1.5,2	0.388	0.344	0.301	0.226
50	12000	0,100	1.5,2	0.377	0.310	0.305	0.181
10	24000	100,1000	1,1.5	0.33	0.28	0.288	0.145
50	12000	100,1000	1,1.5	0.362	0.324	0.312	0.181

Difference in timescales

Coalescent models are usually obtained as approximations of real populations with finite, and often very different, population size N. The difference in time-scale can thus be quite significant. This important aspect of coalescent models is subtly hidden in the actual resulting limiting models. Indeed, the mutation rate μ at the locus under consideration per individual per generation, as was noted by ELDON and WAKELEY (2006), must be scaled in proportion to $1/c_N$, where c_N is the probability that two gene copies, drawn uniformly at random and without replacement from a population of size N, derive from a common parental gene copy in the previous generation. This follows from the famous limit theorem for Cannings models of MÖHLE and SAGITOV (2001). In the usual Wright-Fisher haploid model, $c_N = 1/N$ and θ is proportional to $\mu \cdot N$. In a general Cannings model,

$$c_N = \frac{\mathbb{E}[\nu_1(\nu_1 - 1)]}{N - 1} \tag{S10}$$

where ν_1 is the random number of offspring of individual one (arbitrarily labelled) in any given generation. SCHWEINSBERG (2003) gives a timescale associated with the Beta$(2-\alpha, \alpha)$-coalescent, where $1/c_N$ is proportional to $N^{\alpha-1}$, $1 < \alpha < 2$. Thus, θ is proportional to $\mu \cdot N^{\alpha-1}$ when associated with the Beta$(2-\alpha, \alpha)$-coalescent, as was also noted by BIRKNER and BLATH (2008). As emphasized by ELDON and WAKELEY (2006), this discrepancy in timescales between different coalescent models complicates comparisons of predictions of genetic diversity when only θ is used for mutation rate, since θ associated with the usual Kingman coalescent is often not the same θ as the one associated with a multiple merger coalescent.

Care must also be taken with models of fluctuating population size. In KAJ and KRONE (2003), a time-changed n-coalescent under a general model of variable population size is derived. More precisely, the authors consider a haploid Wright-Fisher model with population size N at generation $r = 0$ and consider a population size process $M_N(r), r \in \mathbb{Z}$ of the form

$$M_N(r) = NX_N(r), \quad r \in \mathbb{Z},$$

that is, $X_N(r)$ describes the 'relative population size' at generation r. Under the assumption that $X_N(\lfloor Nt \rfloor), t \in \mathbb{R}$ converges to something non-degenerate (i=..e. bounded away from 0 and ∞), they get the well-known limiting result that a time-changed Kingman coalescent describes the genealogy, where the infinitesimal coalescence rates are given by $1/\nu(s)$, with

$$\nu(s) = \lim_{N\to\infty} X_N(\lfloor Ns \rfloor). \tag{S11}$$

Our previously discussed exponential growth model corresponds to a growth rate of β/N per generation in the pre-limiting model, and indeed we have

$$\nu(t) = \lim_{N\to\infty} X_N(\lfloor Nt \rfloor) = \lim_{N\to\infty} \left(1 + \frac{\beta}{N}\right)^{Nt} = e^{\beta t}.$$

Thus, the size Nt generations ago is approximately $Ne^{-\beta t}$. In practice, one needs to choose a reference population size (say at present time 0, from data), and then choose a growth function $\nu(s)$, normed in a way that it equals 1 at the current time 0 (one could also take a reference population size at some other time-point, and then rescale the population size to be 1 at that time-point for the computation of the coalescence rates). The chosen reference population size, say N_0, should then be used to convert time into generations. For example, if $S_j^{(n),\nu}$ is the time, measured in coalescent units from the present, at which two out of currently j ancestral lineages coalesce, started from n leaves,

$$G_j^{(n),\nu} = S_j^{(n),\nu} \cdot N_0$$

is the corresponding time in generations at which the two ancestral lineages coalesce.

Minimum-distance statistics

To estimate power under point hypotheses we employed standard ℓ_p minimum-distances measures. Minimal distance estimators are a popular tool to solve the problem of fitting data with a reasonable, though simplified model that cannot be absolutely exact. They

satisfy useful optimality properties (MILLAR, 1984) and are sometimes preferred over a maximum-likelihood approach (see e.g. BERKSON (1980) for a discussion).

The minimum-distance statistics X_p and Y_p we consider are defined by

$$X_p \equiv \left(\sum_{i=1}^{n-1} \left| \xi_i^{(n)} - \mathbb{E}^{\Pi} \left[\xi_i^{(n)} \right] \right|^p \right)^{1/p} \tag{S12}$$

and

$$Y_p \equiv \left(\sum_{i=1}^{n-1} \left| \zeta_i^{(n)} - \varphi_i^{(n),\Pi} \right|^p \right)^{1/p}, \tag{S13}$$

where $\mathbb{E}^{\Pi}\left[\xi_i^{(n)}\right]$ resp. $\varphi_i^{(n),\Pi}$ is obtained under the null hypothesis (Π). Figure S4 shows graphs of X_2 and Y_2 as functions of α when $\mathbb{E}^{\mathtt{E}}\left[\xi_i^{(n)}\right]$ (and $\mathbb{E}^{\mathtt{K}}\left[\xi_i^{(n)}\right]$) is compared to $\mathbb{E}^{\mathtt{B}}\left[\xi_i^{(n)}\right]$ (Figure S4, left), or $\varphi_i^{(n),\mathtt{B}}$ compared to $\varphi_i^{(n),\mathtt{E}}$ (and $\varphi_i^{(n),\mathtt{K}}$; Figure S4, right) for different values of the exponential growth parameter β and sample size n as shown.

To account for the difference in variance of $\xi_i^{(n)}$ resp. $\zeta_i^{(n)}$ one would standardise the deviations $\left(\xi_i^{(n)} - \mathbb{E}^{\Pi}\left[\xi_i^{(n)}\right]\right)$ resp. $\left(\zeta_i^{(n)} - \varphi_i^{(n),\Pi}\right)$ by the standard deviation of $\xi_i^{(n)}$ resp. $\zeta_i^{(n)}$. However, one can only compute the variance $\mathrm{Var}^{\Pi^{(\Lambda)}}\left(\xi_i^{(n)}\right)$ by a recursion, which is $O(n^5)$ (BIRKNER *et al.*, 2013b), and thus only works for quite small n. In addition, we do not have a way of computing $\mathrm{Var}^{\Pi}\left(\zeta_i^{(n)}\right)$.

Distributional convergence results, with knowledge of convergence speed, would allow the application of the MILLAR (1984) machinery, leading to confidence bounds on the minimal distance estimators. Unfortunately, such results are elusive in most cases, and where known, indicate very slow convergence speed.

Power computations

The *power* of a statistical test is the probability of rejecting a false null hypothesis. By the Neyman-Pearson lemma, a likelihood-ratio test has the highest power of all tests of the same size. However, since the distribution of the site-frequency spectrum is unknown, we consider the minimum-distance statistics X_p (S12) and Y_p (S13).

To estimate the power, say, when the null hypothesis is of exponential growth ($\mathtt{E}$) with growth parameter β, and the alternative of a Beta$(2-\alpha, \alpha)$-coalescent ($\mathtt{B}$) with parameter α, the expected values, $\mathbb{E}^{\mathtt{E}}\left[\xi_i^{(n)}\right]$ were estimated under population growth using simulations, making use of the results of Griffiths and Tavaré (1998). The quantiles q_x of the statistic (X_p or Y_p) corresponding to a given size $1-x \in \{0.01, 0.05, 0.1\}$ of the test were then estimated from 10^5 site-frequency spectra simulated under population growth. Since the null hypothesis is assumed to be wrong, site-frequency spectra are now simulated under a Beta$(2-\alpha, \alpha)$-coalescent with a given parameter α, and the power estimated as the fraction of times the computed statistic equalled or exceeded the estimated quantile q_x. In notation, let $\underline{\xi}^{(n),\mathtt{B}}$ denote the site-frequency spectrum associated with the Beta$(2-\alpha, \alpha)$-coalescent resp. exponential population growth ($\underline{\xi}^{(n),\mathtt{E}}$), ie. when $\underline{\xi}^{(n)}$ is simulated under the model indicated. We estimate the quantiles q_x of the statistic

$$X_p^{\mathtt{E},\mathtt{E}} = \left(\sum_j \left|\xi_j^{(n),\mathtt{E}} - \mathbb{E}^{\mathtt{E}}\left[\xi_j^{(n)}\right]\right|^p\right)^{1/p}.$$

by simulating values of $\xi_j^{(n),\mathtt{E}}$. Define the statistic $X_p^{\mathtt{B},\mathtt{E}}$ by

$$X_p^{\mathtt{B},\mathtt{E}} = \left(\sum_j \left|\xi_j^{(n),\mathtt{B}} - \mathbb{E}^{\mathtt{E}}\left[\xi_j^{(n)}\right]\right|^p\right)^{1/p}.$$

which refers to population growth being the null hypothesis, and the Beta$(2-\alpha, \alpha)$-coalescent being the alternative. The power $\mathbb{P}\left(X_p^{(\Lambda,\mathtt{E})} \geq q_x\right)$ of the statistic $X_p^{(\Lambda,\mathtt{E})}$ is estimated by

$$\mathbb{P}\left(X_p^{(\Lambda,\mathtt{E})} \geq q_x\right) \doteq \frac{1}{R}\sum_{b=1}^{R} \mathbf{1}_{\left(\hat{X}_p^{(b,\Lambda,\mathtt{E})} \geq q_x\right)},$$

where $\hat{X}_p^{(b,\Lambda,\mathtt{E})}$ denotes the value of the statistic $X_p^{(\Lambda,\mathtt{E})}$ computed for the b-th replicate and R the number of replicates. In a similar way, of course, one can estimate power when the null hypothesis is of a Beta$(2-\alpha, \alpha)$-coalescent with alternative being exponential growth.

Figure S5: Estimation of $\mathbb{P}(Y_2 \geq q_x)$ - the power for the Y_2 norm (S13) for $x \in \{0.9, 0.95, 0.99\}$ as shown in the legend - as a function of sample size $n \in \{10, 25, 50, 75, 100, 150, 200, 250, 300, 500, 750, 1000\}$. The alternative is exponential growth with β as shown, and the null of a Beta-coalescent with $\alpha \in \{5/4, 3/2, 7/4\}$ and total number of segregating sites s as shown, with mutation rate μ per generation estimated from s each time. Expected values, quantiles, and power estimates based on 10^5 iterations each.

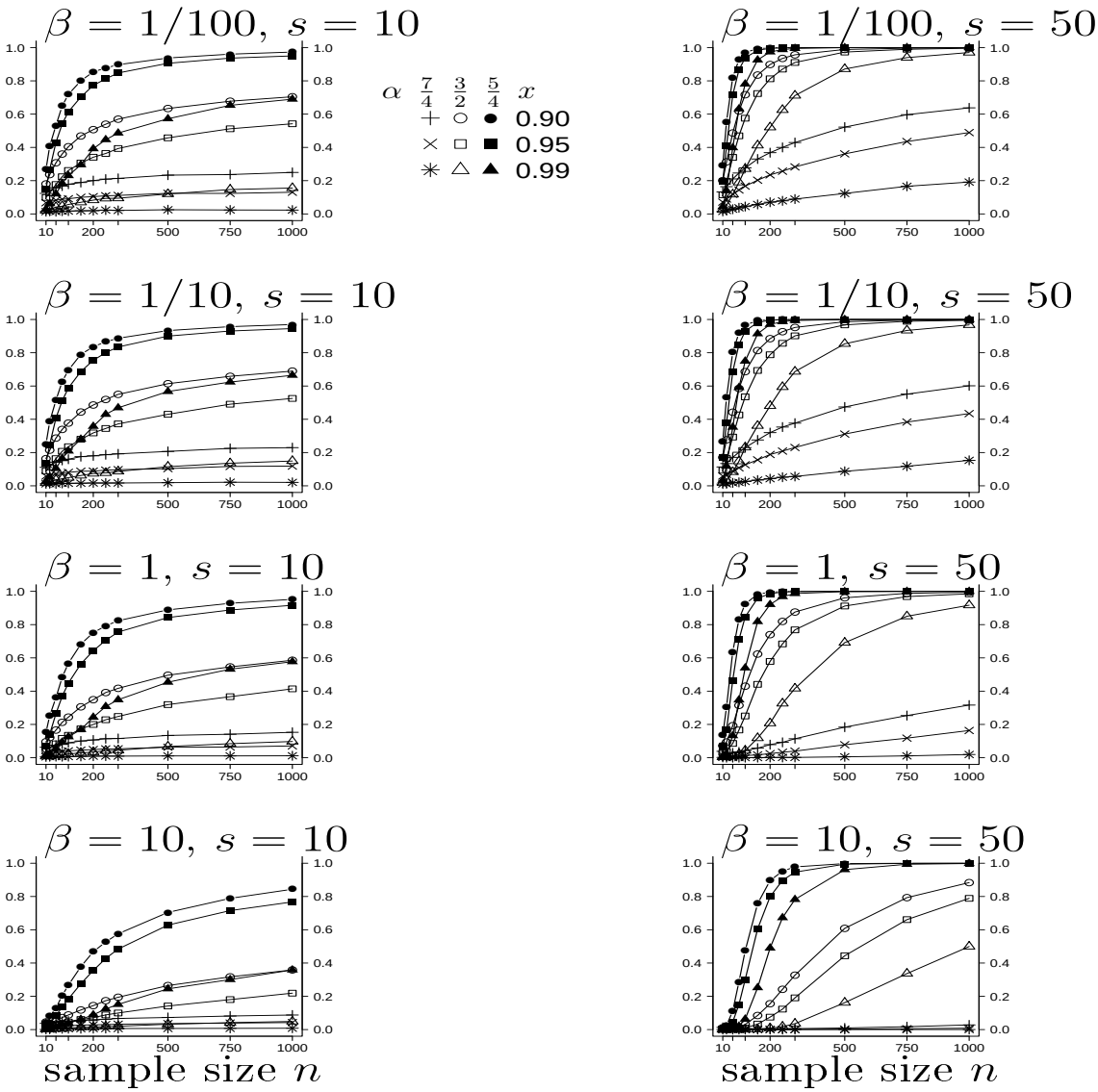


Figure S6: Estimation of $\mathbb{P}\left(X_2^{(-1)} \geq q_x\right)$ - the power for the X_2 norm (S12) excluding the singletons for x as shown in the legend - as a function of sample size $n \in \{10, 25, 50, 75, 100, 150, 200, 250, 300, 500, 750, 1000\}$ and time computed in generations with current population size $N_0 = 10^4$. The alternative hypothesis is of an exponential growth with β as shown, and the null of a Beta-coalescent with α and total number of segregating sites s as shown, with μ estimated from s each time. Expected values, quantiles, and power estimated based on 10^5 iterations each.

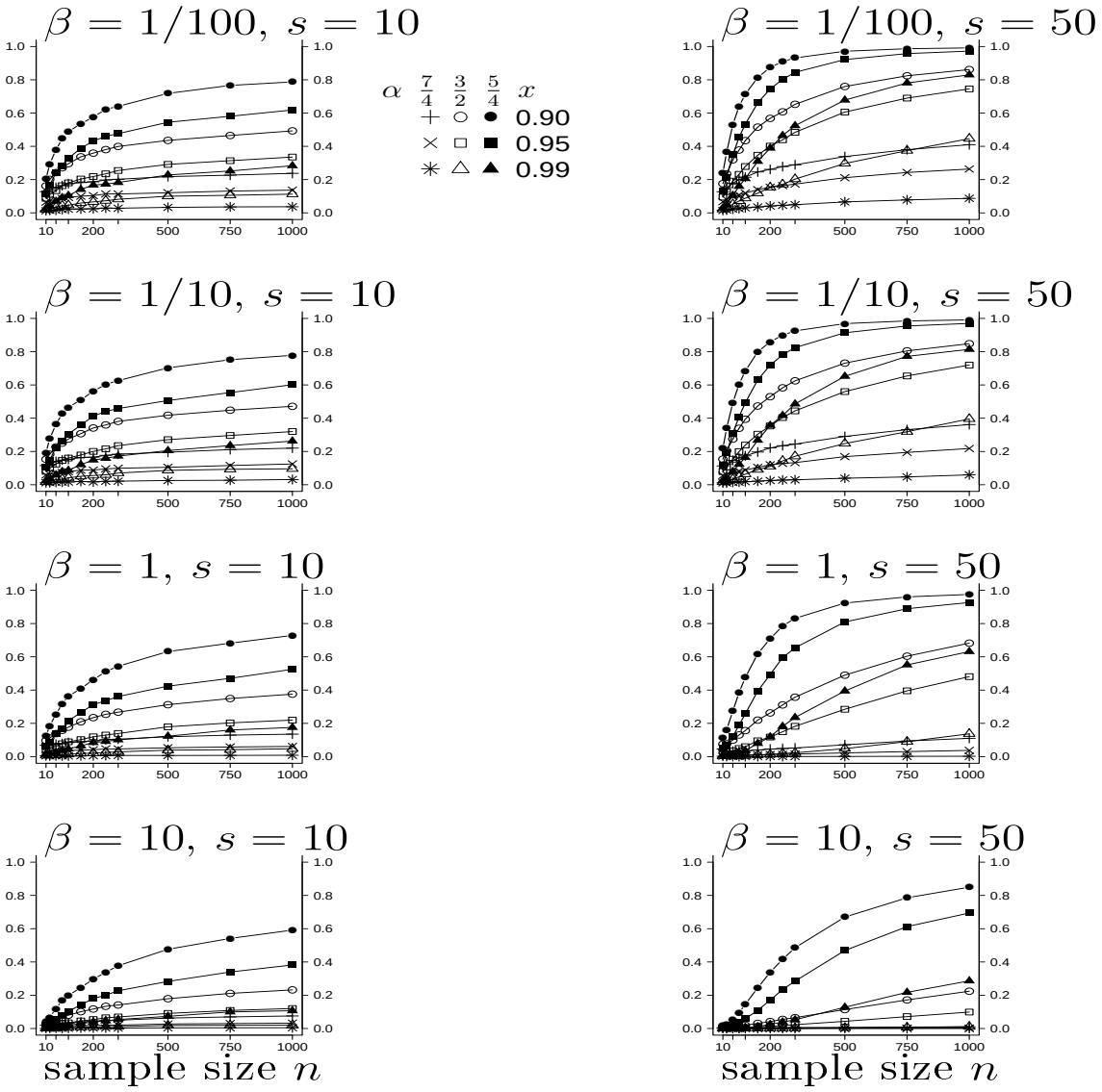
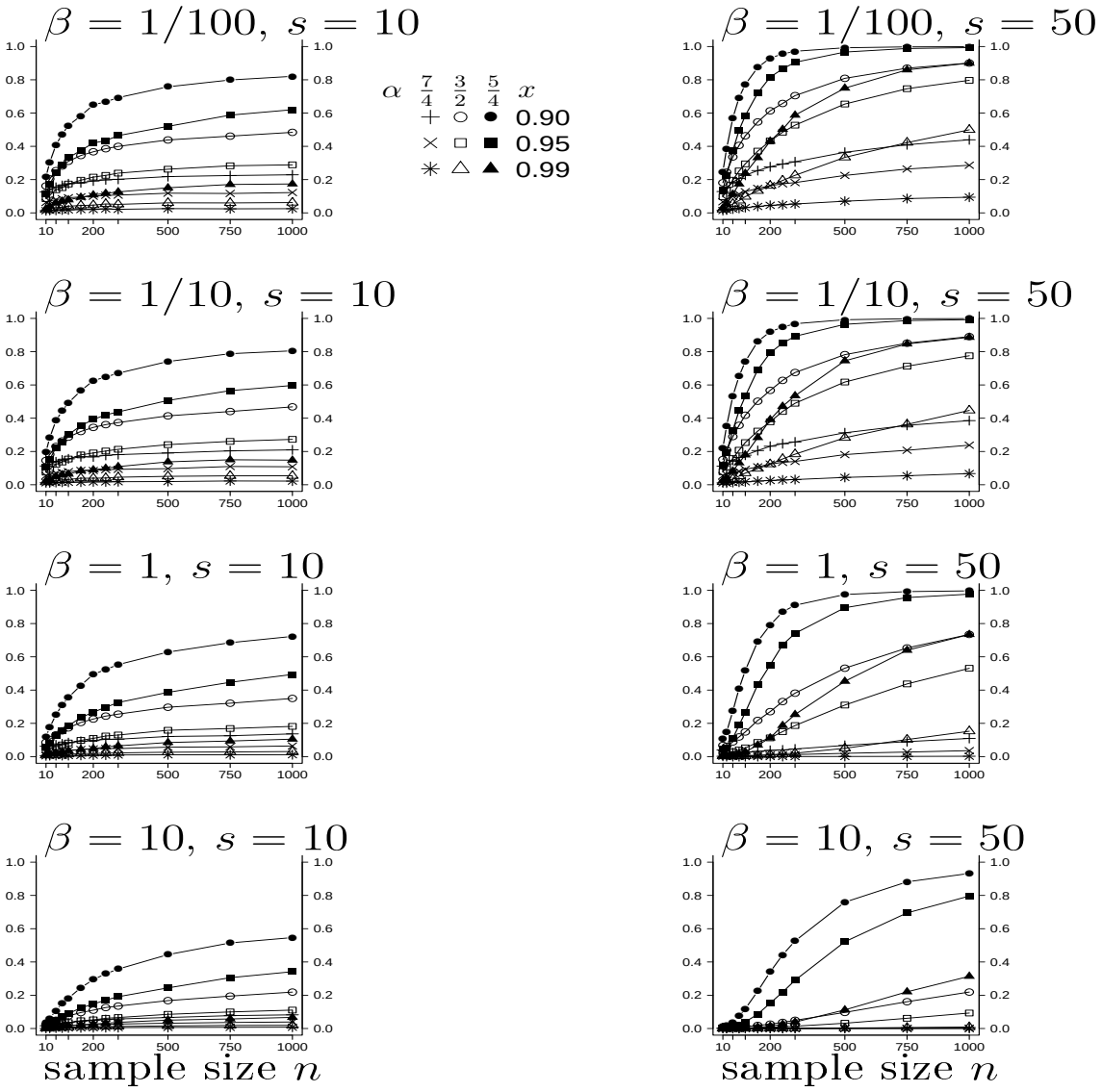


Figure S7: Estimation of $\mathbb{P}\left(Y_2^{(-1)} \geq q_x\right)$ - the power for the Y_2 norm (S13) excluding the singletons for x as shown in the legend - as a function of sample size $n \in \{10, 25, 50, 75, 100, 150, 200, 250, 300, 500, 750, 1000\}$ and time computed in generations with current population size $N_0 = 10^4$. The alternative hypothesis is of an exponential growth with β as shown, and the null of a Beta-coalescent with α and total number of segregating sites s as shown, with μ estimated from s each time. Expected values, quantiles, and power estimated based on 10^5 iterations each.



Deriving the expected SFS under variable population size

The computations outlined below for the reader's convenience are essentially contained in Polanski *et al.* (2003).

For notational simplicity, we drop the index n for the sample size and the index Π resp. ν for the underlying coalescent model resp. the demographic history. Consider a time-inhomogeneous n-coalescent where each pair of lines present at time t merges at rate $\nu(t)$, and recall that T_k, $k = n, n-1, \ldots, 2$ denote the length of the time interval while there are exactly k lineages, and let $S_j := T_n + T_{n-1} + \cdots + T_j$, $j = n, \ldots, 2$, the time point when the number of lines jumps from j to $j-1$, with and $S_{n+1} := 0$. Then, from (Griffiths and Tavaré, 1998, (2.4)), we have

$$\mathbb{P}(T_k \in (t, t+dt) | S_{k+1} = s) = \binom{k}{2} \nu(s+t) \exp\left(-\binom{k}{2} \int_s^{s+t} \nu(u)\, du \right) dt \tag{S14}$$

so the joint density of $T_m, \ldots, T_n$ $(2 \le m \le n)$ is

$$\prod_{k=m}^{n} \binom{k}{2} \nu\Big(\sum_{k \le j \le n} t_j \Big) \times \exp\left(- \sum_{k=m}^{n} \binom{k}{2} \int_{\sum_{k+1 \le j \le n} t_j}^{\sum_{k \le j \le n} t_j} \nu(u)\, du \right). \tag{S15}$$

Hence, the joint density of the $S_m, \ldots, S_n$ is given by

$$\begin{aligned} &\prod_{k=m}^{n} \binom{k}{2} \nu(s_k) \times \exp\left(- \sum_{k=m}^{n} \binom{k}{2} \int_{s_{k+1}}^{s_k} \nu(u)\, du \right) \\ &\qquad = \binom{m}{2} \nu(s_m) e^{-\binom{m}{2} F(s_m)} \times \prod_{k=m+1}^{n} \binom{k}{2} \nu(s_k) e^{-(k-1)F(s_k)}, \end{aligned} \tag{S16}$$

$(0 = s_{n+1} < s_n < s_{n-1} < \cdots < s_m,\ 2 \le m \le n)$, with $F(t) := \int_0^t \nu(u)\, du$ where we used that $\binom{k}{2} = \sum_{j=2}^{k} (j-1)$ to express

$$\begin{aligned} \sum_{k=m}^{n} \binom{k}{2} \int_{s_{k+1}}^{s_k} \nu(u)\, du &= \sum_{j=2}^{n} \sum_{k=j \vee m}^{n} (j-1) \int_{s_{k+1}}^{s_k} \nu(u)\, du \\ &= \binom{m}{2} \int_0^{s_m} \nu(u)\, du + \sum_{j=m+1}^{n} (j-1) \int_0^{s_j} \nu(u)\, du. \end{aligned}$$

The marginal density of S_m at fixed s_m can e.g. be found by integrating out over $s_n < s_{n-1} < \cdots <$

s_{m+1} ($< s_m$), it can be expressed as a "generalised mixture" of densities[1], as follows:

$$f_{S_m}(s_m) = \sum_{j=m}^{n} c_m^{(j,n)} \nu(s_m) \binom{j}{2} e^{-\binom{j}{2} F(s_m)} \tag{S17}$$

where the coefficients $c_m^{(\ell,n)}$ can be computed (backwards) recursively: Let $c_n^{(n,n)} = 1$, and

$$c_m^{(j,n)} = -c_{m+1}^{(j,n)} \frac{\binom{m}{2}}{\binom{j}{2} - \binom{m}{2}}, \qquad j = m+1, m+2, \ldots, n, \; 2 \le m \le n-1, \tag{S18}$$

$$c_m^{(m,n)} = \sum_{j=m+1}^{n} c_{m+1}^{(j,n)} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{m}{2}} = 1 - \sum_{j=m+1}^{n} c_m^{(j,n)}, \qquad 2 \le m \le n-1, \tag{S19}$$

(and $c_m^{(\ell,n)} = 0$ for $\ell < m$). An explicit formula is

$$c_m^{(j,n)} = \prod_{\substack{m \le i \le n \\ i \ne j}} \frac{\binom{i}{2}}{\binom{i}{2} - \binom{j}{2}} = (-1)^{j-m} \frac{(2j-1)m}{j(j-1)} \frac{\binom{n}{j}\binom{j+m-2}{j}\binom{j}{m}}{\binom{n+j-1}{j}}. \tag{S20}$$

We check (S17) by induction: $f_{S_n}(s_n) = \binom{n}{2} \nu(s_n) e^{-\binom{n}{2} F(s_n)}$, using the induction hypothesis we find from (S14)

$$\begin{aligned}
f_{S_m}(s_m) &= \int_0^{s_m} \binom{m}{2} \nu(s_m) e^{-\binom{m}{2}(F(s_m) - F(s_{m+1}))} \times f_{S_{m+1}}(s_{m+1})\, ds_{m+1} \\
&= \binom{m}{2} \nu(s_m) e^{-\binom{m}{2} F(s_m)} \sum_{j=m+1}^{n} c_{m+1}^{(j,n)} \binom{j}{2} \int_0^{s_m} \nu(s_{m+1}) \exp\Big(- \big(\tbinom{j}{2} - \tbinom{m}{2}\big) F(s_{m+1}) \Big) ds_{m+1} \\
&= \binom{m}{2} \nu(s_m) e^{-\binom{m}{2} F(s_m)} \sum_{j=m+1}^{n} c_{m+1}^{(j,n)} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{m}{2}} \Big(1 - \exp\Big(- \big(\tbinom{j}{2} - \tbinom{m}{2}\big) F(s_m) \Big) \Big) \\
&= \nu(s_m) \binom{m}{2} e^{-\binom{m}{2} F(s_m)} \sum_{j=m+1}^{n} c_{m+1}^{(j,n)} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{m}{2}} \\
&\quad - \sum_{j=m+1}^{n} c_{m+1}^{(j,n)} \frac{\binom{m}{2}}{\binom{j}{2} - \binom{m}{2}} \nu(s_m) \binom{j}{2} e^{-\binom{j}{2} F(s_m)}
\end{aligned}$$

[1] (S17) is a generalisation of the well known fact that the density of a convolution of exponentials with different rates can be written as a linear combination of exponential densities (set $\nu(\cdot) \equiv 1$ and thus $F(t) = t$ in (S17)): $X_1, \ldots, X_k$ indep., $X_i \sim \mathrm{Exp}(\lambda_i)$ (and the λ_i pairwise different, say $\lambda_1 > \lambda_2 > \cdots > \lambda_k$), then the density of $X_1 + \cdots + X_k$ is $\sum_{i=1}^k a_i \lambda_i e^{-\lambda_i x}$ with $a_i = \prod_{j \ne i}^k \lambda_j / (\lambda_j - \lambda_i)$ which can also be easily checked by considering the characteristic functions.

noting that

$$\frac{d}{ds}\big(-a^{-1}e^{-aF(s)}\big) = \nu(s)e^{-aF(s)}$$

for $a \in \mathbb{R}$. For the second equality in (S19) note (either inductively or from the fact that (S17) must be a probability density) that

$$\sum_{j=m}^{n} c_m^{(j,n)} = 1 \qquad \text{for } 2 \le m \le n.$$

The explicit form (S3) can be checked by verifying inductively that the expression on the right solves (S18)–(S19) (with the interpretation that the empty product = 1). While checking (S18) is straightforward, for checking (S19) one observes that for pairwise different $\lambda_1, \dots, \lambda_k \neq 0$ and any $z \in \mathbb{C}$ we have

$$\prod_{\ell=1}^{k} \frac{\lambda_\ell}{\lambda_\ell - z} = \sum_{j=1}^{k} \frac{\lambda_j}{\lambda_j - z} \prod_{\substack{1 \le i \le k \\ i \neq j}} \frac{\lambda_i}{\lambda_i - \lambda_j}\,; \qquad \text{put } z = 0 \text{ to see that } \quad \sum_{j=1}^{k} \prod_{\substack{1 \le i \le k \\ i \neq j}} \frac{\lambda_i}{\lambda_i - \lambda_j} = 1 \tag{S21}$$

(by partial fraction expansion, observe that both sides of the left equation are meromorphic functions of z with k simple poles at $\lambda_1, \dots, \lambda_k$ whose Laurent coeffients agree).

To check the second equality in (S3) write for fixed $j \in \{m, \dots, n\}$

$$\begin{aligned}
\prod_{\substack{m \le i \le n \\ i \neq j}} \frac{\binom{i}{2}}{\binom{i}{2} - \binom{j}{2}} &= \prod_{\substack{m \le i \le n \\ i \neq j}} \frac{i(i-1)}{(i-j)(i+j-1)} \\
&= \frac{2j-1}{j(j-1)} \prod_{i=m}^{n} \frac{i(i-1)}{i+j-1} \times \left(\prod_{i=m}^{j-1} (i-j) \times \prod_{i=j+1}^{n} (i-j) \right)^{-1} \\
&= (-1)^{j-m} \frac{2j-1}{j(j-1)} \frac{n!}{(m-1)!} \frac{(n-1)!}{(m-2)!} \frac{(m+j-2)!}{(n+j-1)!} \frac{1}{(j-m)!\,(n-j)!} \\
&= (-1)^{j-m} \frac{(2j-1)m}{j(j-1)} \frac{n!}{j!(n-j)!} \frac{(j+m-2)!}{(m-2)!j!} \frac{j!}{m!(j-m)!} \frac{j!(n-1)!}{(n+j-1)!} \\
&= (-1)^{j-m} \frac{(2j-1)m}{j(j-1)} \frac{\binom{n}{j}\binom{j+m-2}{j}\binom{j}{m}}{\binom{n+j-1}{j}}.
\end{aligned}$$

The tail of S_m is given by (assuming $F(\infty) = \infty$ so that all S_m are a.s. finite)

$$\mathbb{P}(S_m > s) = \int_s^\infty f_{S_m}(u)\, du = \sum_{j=m}^{n} c_m^{(j)} e^{-\binom{j}{2}F(s)}.$$

Recalling

$$a_j = \int_0^\infty e^{-\binom{j}{2}F(s)}\, ds,$$

(assuming that F grows sufficiently fast at ∞) then gives the desired result:

$$\mathbb{E}[S_m] = \int_0^\infty \mathbb{P}(S_m > s)\, ds = \sum_{j=m}^{n} c_m^{(j,n)} a_j\,,$$

and

$$\mathbb{E}[T_k] = \mathbb{E}[S_k - S_{k+1}] = c_k^{(k,n)} a_k + \sum_{j=k+1}^{n} \big(c_k^{(j,n)} - c_{k+1}^{(j,n)}\big) a_j\,.$$

For the case of exponential growth, i.e. $\nu(t) = e^{\beta t}$ and

$$F(t) = \int_0^t \nu(u)\, du = \beta^{-1}\left(e^{\beta t} - 1\right), \tag{S22}$$

we get for $c > 0$

$$\int_0^\infty e^{-cF(s)}\, ds = e^{c/\beta} \int_0^\infty \exp\Big(-\tfrac{c}{\beta} e^{\beta s}\Big)\, ds = \frac{1}{\beta} e^{c/\beta} \int_{c/\beta}^\infty e^{-u} \frac{du}{u} = \frac{1}{\beta} e^{c/\beta} E_1(c/\beta) \tag{S23}$$

where we substituted $\frac{c}{\beta} e^{\beta s} = u$ and

$$E_1(t) = \int_t^\infty \frac{e^{-x}}{x}\, dx = \int_1^\infty \frac{e^{-tx}}{x}\, dx$$

is an exponential integral function (e.g. (Abramowitz and Stegun, 1964, 5.1.1)). In particular,

we get

$$a_j = \frac{1}{\beta} \exp\left(\beta^{-1}\tbinom{j}{2}\right) E_1\left(\beta^{-1}\tbinom{j}{2}\right).. \tag{S24}$$

For the case of algebraic growth consider $\nu(t) = t^\gamma$ for some $\gamma > 0$. Then

$$F(t) = \int_0^t s^\gamma \, ds = \frac{1}{\gamma+1} t^{\gamma+1}$$

and for $c > 0$,

$$\begin{aligned} \int_0^\infty e^{-cF(s)} \, ds &= \int_0^\infty e^{-c(\gamma+1)^{-1} s^{\gamma+1}} \, ds = c^{-1/(\gamma+1)} \gamma^{-\gamma/(\gamma+1)} \int_0^\infty e^{-u} u^{-\gamma/(\gamma+1)} \, du \\ &= c^{-1/(\gamma+1)} \gamma^{-\gamma/(\gamma+1)} \Gamma\big(1/(\gamma+1)\big) \end{aligned} \tag{S25}$$

where we substituted $u = ct^{\gamma+1}/(\gamma+1)$, hence $du/dt = ct^\gamma = c^{1/(\gamma+1)} \gamma^{\gamma/(\gamma+1)} u^{\gamma/(\gamma+1)}$. In particular, we obtain

$$a_j = \frac{\Gamma\big(1/(\gamma+1)\big)}{\gamma^{\gamma/(\gamma+1)}} \binom{j}{2}^{-1/(\gamma+1)}. \tag{S26}$$

ABC analysis of the cytochrome *b* mtDNA data of Árnason (2004)

To investigate which model class fits better to the data, we use the ABC model comparison approach given the (lumped) nfSFS of the observed mitochondrial locus. The growth model class is specified by an uniform prior on $\{0, 10, 20, \ldots, 1000\}$ for the class of growth models and the class of Beta n-coalescents by an uniform prior on $\{1, 1.05, \ldots, 2\}$. Due to the numeric difficulties of evaluating the exact expected tree length for the used models, we approximate $E^\Pi\left[B^{(n)}\right]$ for the prior mutation rate $\frac{2s}{E^\Pi\left[B^{(n)}\right]}$ corresponding to a chosen α or β by the mean value from 10,000 simulations. We use a tolerance level of 0.005 and perform $n_{reps} = 200,000$ simulations for each model class. See Table S4 for the approximated Bayes factors $\varrho^{\mathtt{E/B}}$ for the model comparison of the growth model and the Beta n-coalescent

model using different lumps of the nfSFS as summary statistics. The observed data fits slightly better to the growth model, but not so much better that we could discard the Beta n-coalescents as possible genealogy models for this locus. JEFFREYS (1961) suggested

Table S4: Approximated Bayes factors given the Atlantic cod mtDNA data

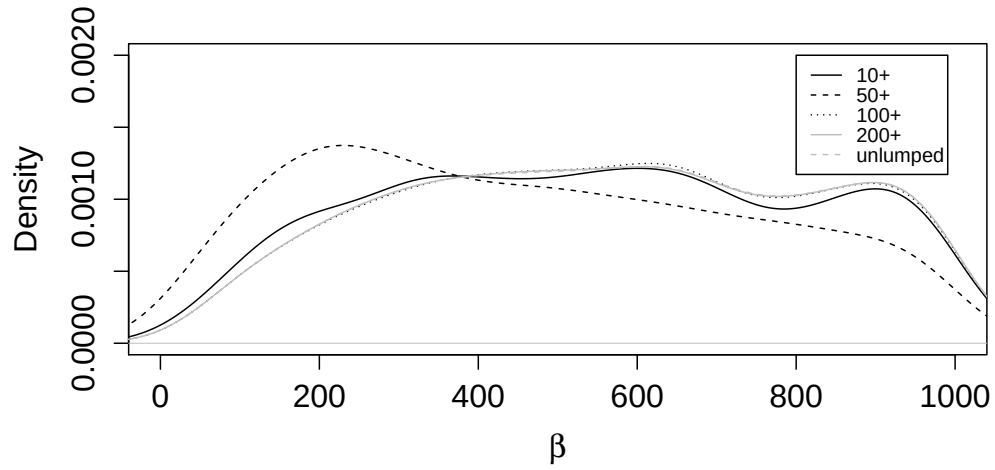
lumping number	10+	50+	100+	200+	no
$\varrho^{\mathtt{E/B}}$	6.435	1.74	2.378	3.264	3.175

interpreting Bayes factors according to the $\log_{10}$ scale. Lumping at 10 (Table S4) then gives 'substantial' $(1/2 < \log_{10}(\varrho^{\mathtt{E/B}}) < 1)$ evidence against the Beta$(2-\alpha, \alpha)$-coalescent in favor of exponential growth. Using KASS and RAFTERY (1995) suggestion of considering Bayes factors on $2\log_e$ scale gives 'positive' $(2 < 2\log_e(\varrho^{\mathtt{E/B}}) < 6)$ evidence in favor of exponential growth, based on lumping at 10.

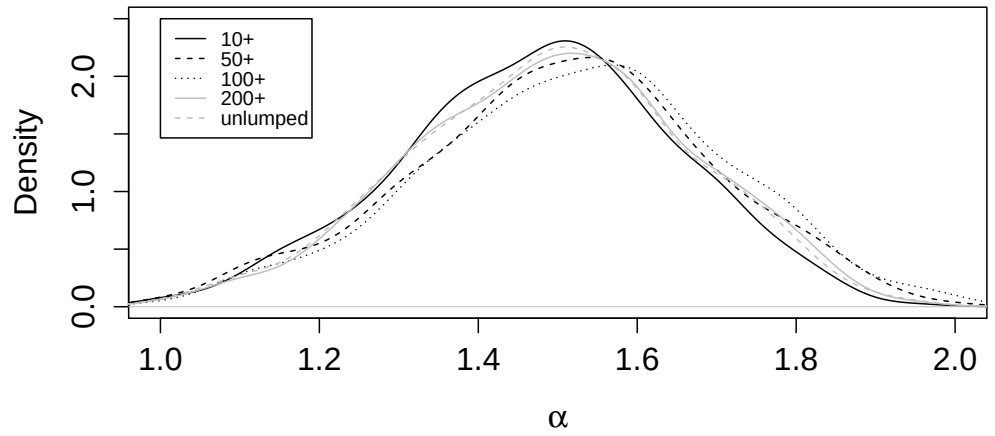
Additionally to the ABC model comparison, we also evaluate which parameters fit best to the observed nfSFS at the mitochondrial locus. For each model class used, we record the prior parameters from the 0.5% of the $n_{reps} = 200,000$ simulations that have the smallest ℓ^2 distance to the observed nfSFS (summary statistics). This gives an approximate sample of the posterior distribution of $\pi(\alpha|$ observed $\underline{\zeta}^{(n)})$ resp. $\pi(\beta|$ observed $\underline{\zeta}^{(n)})$. Analogously, we also used the lumped nSFS as summary statistics. Figure S8 shows the posterior distributions for different lumping numbers.

Figure S8: Approximate ABC sample from ABC fitting of the **A** growth and **B** Beta n-coalescent model classes to the observed nfSFS in the Atlantic cod data. Denote by α the Beta n-coalescent parameter, β the growth rate. Priors were uniform on both sets.

A



B



ABC quality control for the Árnason (2004) data

We follow the recommendation from the R package abc (Csilléry *et al.*, 2012) and perform three checks of quality for the presented ABC approach. We focus on the lumping which gives the clearest distinction, namely the lumping of all classes with mutation counts 10 or higher (class 10+). All checks are performed using the R package abc

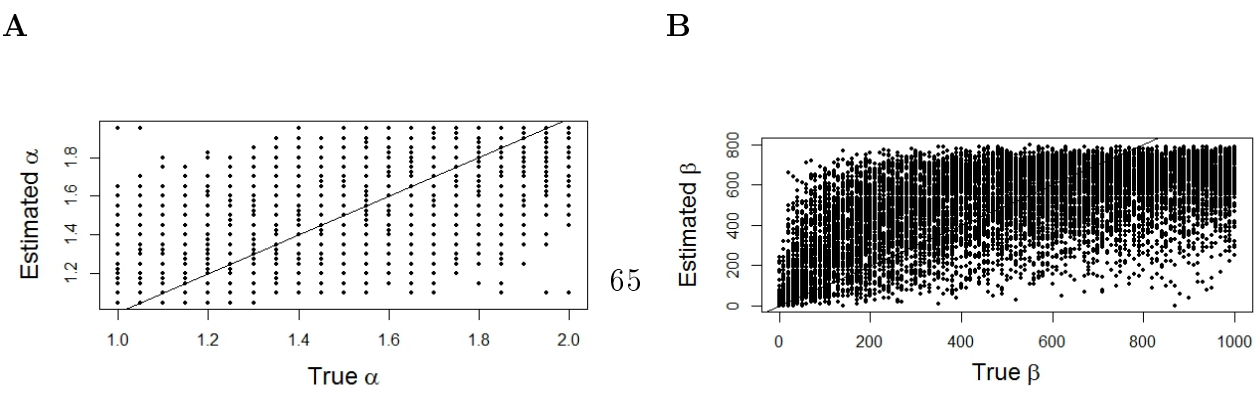
To assess the general ability to distinguish between the two model classes in the setting (i.e., number of observed mutations and sample size) given by the Atlantic cod mtDNA data from Árnason (2004), we again employ a leave-one-out cross-validation as described in Methods. See Table S5 for the results.

Table S5: Mean posterior probabilities and misclassification probabilities for the ABC model comparison for tolerance $x = 0.005$ and $n = 1278$. We use the number $s = 39$ of observed mutations to estimate the mutation rate via Watterson's estimator and use the nfSFS (10+ lumped) as summary statistics. We use $n_{cv} = 12,000$ cross validations.

$\mathbb{E}_{\mathtt{B}}\left[\pi(\mathtt{E}\vert\underline{\zeta})\right]$	$\mathbb{E}_{\mathtt{E}}\left[\pi(\mathtt{B}\vert\underline{\zeta})\right]$	$\mathbb{E}_{\mathtt{B}}\left[\pi(\varrho^{\mathtt{E/B}} > 1\vert\underline{\zeta})\right]$	$\mathbb{E}_{\mathtt{E}}\left[\pi(\varrho^{\mathtt{E/B}} < 1\vert\underline{\zeta})\right]$
0.283	0.238	0.232	0.13

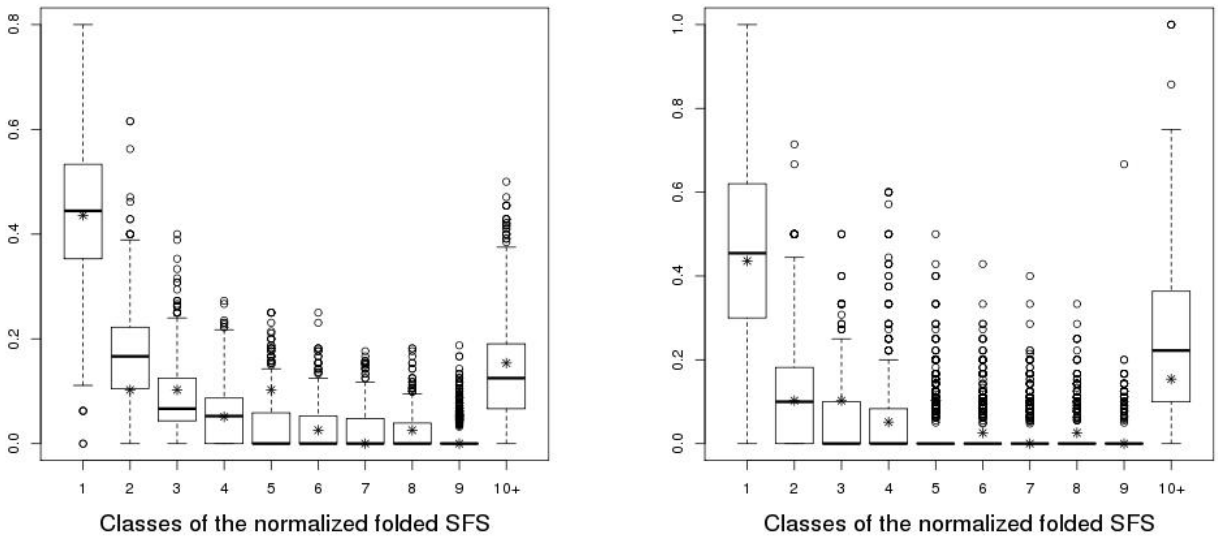
To assess the quality to distinguish the parameters within one model class, we again use leave-one-out cross-validations ($n_{cv} = 12,000$). The parameter of each simulation chosen for cross-validation is estimated as the median of the 0.5% of simulations with the smallest ℓ^2 distance to the chosen simulation. Figure S9 shows the resulting scatter plots of the parameters of the chosen simulations and the corresponding estimations.

Figure S9: Scatter plots of estimated vs. true parameters of $n_{cv} = 12000$ cross-validated simulations in the **A** Beta coalescent model class **B** growth model class



To see whether the posterior distributions given the cod mtDNA data from ÁRNASON (2004) define models under which the observed data is reproducible, we perfomed posterior predictive checks by simulating the 10+ lumped nfSFS under the posterior distribution (i.e., simulating once from each parameter set of each of the 1,000 accepted simulations) for each model class and compare these with the nfSFS observed. See Figure S10 for the results.

Figure S10: Posterior predictive checks with 1,000 simulations of the nfSFS under the approximate posterior distributions given the cod data from ÁRNASON (2004) for the **A** Beta coalescent model class **B** growth model class. Asterisks denote the observed values in the data.

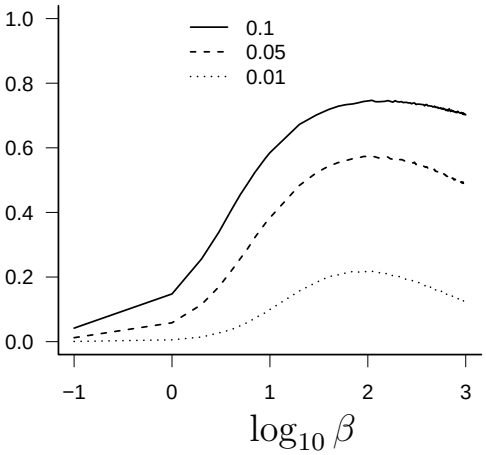


The quality checks reveal that we can't distinguish well within each model class, but moderately between model classes. The posterior predictive checks reveal that both model classes can produce the observed values, thus including possible (though not necessarily well-fitting) models for the data at hand.

Estimation of power for $n = 100$

Figure S11: Estimate of power as a function of $\log_{10}\beta$ for $\beta \in \{0, 1, 2, \ldots, 9, 10, 20, \ldots, 1000\}$ when the Beta$(2-\alpha, \alpha)$-coalescentis the null hypothesis, and the test statistic is $\sup\{\ell(\vartheta, \xi^{(n)}), \vartheta \in \Theta_0\} - \sup\{\ell(\vartheta, \xi^{(n)}), \vartheta \in \Theta_1\}$ (5) , with $\ell(\vartheta, \xi^{(n)})$ the log of the Poisson likelihood function (9) (no lumping). Values at $\log_{10}\beta = -1$ correspond to the Kingman coalescent ($\beta = 0$). In **A**, 10^5 replicates; in **B**, 10^6 replicates.

A $n = 100,\ s = 50$



B $n = 100,\ s = 300$

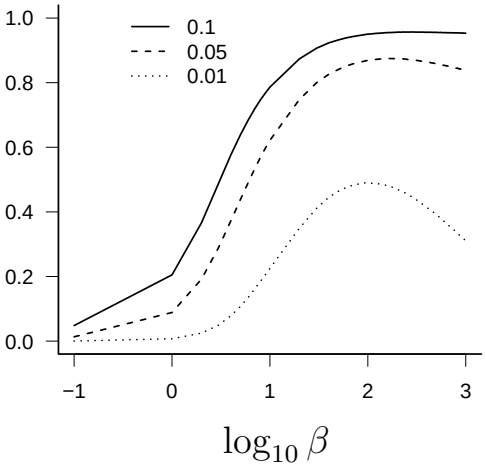


Figure S12: Estimate of power as a function of α for $\alpha \in [1, 2]$ when the Beta$(2-\alpha, \alpha)$-coalescentis the null hypothesis, and the test statistic is $\sup\{\ell(\vartheta, \xi^{(n)}), \vartheta \in \Theta_0\} - \sup\{\ell(\vartheta, \xi^{(n)}), \vartheta \in \Theta_1\}$ (5) , with $\ell(\vartheta, \xi^{(n)})$ the log of the Poisson likelihood function (9) (no lumping). Values at $\alpha = 2$ correspond to the Kingman coalescent; 10^6 replicates for quantiles and power estimates.