VIRUS EVOLUTION

Virus Evolution, 2019, 5(2): vez031

doi: 10.1093/ve/vez031
Research article

The multifurcating skyline plot

Patrick Hoscheit[1,*,†] and Oliver G. Pybus[2]

[1]MaIAGE, INRA, Université Paris-Saclay, Domaine de Vilvert, Jouy-en-Josas 78350, France and [2]Department of Zoology, University of Oxford, Peter Medawar Building, South Parks Road, Oxford OX1 3SY, UK

*Corresponding author: E-mail: patrick.hoscheit@inra.fr

†http://orcid.org/0000-0002-8001-348X

Abstract

A variety of methods based on coalescent theory have been developed to infer demographic history from gene sequences sampled from natural populations. The 'skyline plot' and related approaches are commonly employed as flexible prior distributions for phylogenetic trees in the Bayesian analysis of pathogen gene sequences. In this work we extend the classic and generalized skyline plot methods to phylogenies that contain one or more multifurcations (i.e. hard polytomies). We use the theory of Λ-coalescents (specifically, $Beta(2-\alpha,\alpha)$-coalescents) to develop the 'multifurcating skyline plot', which estimates a piecewise constant function of effective population size through time, conditional on a time-scaled multifurcating phylogeny. We implement a smoothing procedure and extend the method to serially sampled (heterochronous) data, but we do not address here the problem of estimating trees with multifurcations from gene sequence alignments. We validate our estimator on simulated data using maximum likelihood and find that parameters of the $Beta(2-\alpha,\alpha)$ -coalescent process can be estimated accurately. Furthermore, we apply the multifurcating skyline plot to simulated trees generated by tracking transmissions in an individual-based model of epidemic superspreading. We find that high levels of superspreading are consistent with the high-variance assumptions underlying Λ-coalescents and that the estimated parameters of the Λ-coalescent model contain information about the degree of superspreading.

Key words: phylogenetics; coalescent; multifurcating; maximum likelihood; phylodynamics.

1. Introduction

The field of *phylodynamics* is concerned with the study of how processes acting at the population-level shape the genetic diversity of gene or genome sequences sampled from natural populations. Phylodynamic methods are frequently applied to pathogen populations and used to test hypotheses concerning the epidemiology and transmission of infectious disease (e.g. Pybus and Rambaut 2009). Viruses in particular have been the subject of great attention, since their high mutation rates rapidly generate genetic diversity, even on short time scales, and because increasingly large numbers of virus genome sequences from viral epidemics are available for analysis (e.g. Biek et al. 2015). Phylodynamic approaches are also used in other fields, such as macroevolution, anthropology, and ancient DNA research (e.g. Drummond et al. 2003), in order to understand how dynamical processes gave rise to the patterns of ancestry and diversity observed in biological systems.

Phylodynamic analysis of sampled gene sequences relies crucially on 'tree-generating models', which describe how phylogenies, genealogies or trees (we use these terms interchangeably) are related to the population dynamic processes that generated them. Among these models, coalescent approaches are widely used because they provide a mathematically simple framework that is capable of relating the demography of a viral population to its sample genealogy. Mathematically, coalescent theory relies on asymptotic properties of Wright–Fisher reproduction models that represent large, constant-sized populations. The distribution of sample genealogies from such populations is described by the so-called Kingman coalescent process (Kingman 1982). The Kingman coalescent has been shown to describe the

genealogy of many models in population genetics and has been extended to incorporate a number of biological processes, including population size change (Griffiths and Tavare 1994), selection (Kaplan, Darden, and Hudson 1988), recombination (Hudson and Kaplan 1988), longitudinal sampling (Rodrigo et al. 1999), and population structure (Takahata 1988).

The Kingman coalescent has been used to develop a variety of statistical methods that aim to infer the history of population size from an observed sample phylogeny. One such approach that is commonly used is the 'skyline plot' and related methods (Ho and Shapiro 2011), which model population size change as a piecewise constant function through time. In this paper, we extend and generalize the skyline plot family of methods beyond the Kingman coalescent.

The standard Kingman coalescent (Kingman 1982) describes the genealogy of n individuals sampled at random from a population of size $N \gg n$, using a bifurcating ultrametric tree T_n with n leaves (tree tips). In this paper, we will consider genealogies obtained by sampling from large populations whose sizes vary over time. In contrast to standard Kingman coalescent theory, we will consider populations with high variance in the number of offspring per individual (sometimes called the offspring distribution) which may lead to multifurcations (i.e. nodes with degree >3) in the sample genealogy that can no longer be ignored. To achieve this we consider a more general class of models called Λ-coalescent processes, a family of random trees discovered by Sagitov (1999) and fully described by Pitman (1999). In contrast to the Kingman coalescent, under which trees are binary and strictly bifurcating, Λ-coalescent trees can contain multifurcations. This has led to their application to genetic data from populations undergoing high-variance reproduction (sometimes called *sweepstakes reproduction*), such as the Atlantic cod *Gadus morhua* (Birkner, Blath, and Eldon 2013) or Pacific oysters *Crassostrea gigas* (Sargsyan and Wakeley 2008). Beyond high-variance reproduction, many other phenomena can lead to multifurcations in the ancestral process of a sample, such as repeated selective sweeps (Durrett and Schweinsberg 2004), strong selective pressure (Neher and Hallatschek 2013), or oversampling (Bhaskar, Clark, and Song 2014). Using classical methods to analyze such datasets can lead to systematic biases (Hallatschek 2018; Sackman, Harris, and Jensen 2019).

Most work in this area has been focused on finding statistical signatures of multiple-merger coalescences, and distinguishing them from confounding factors such as population growth (Eldon et al. 2015; Koskela 2018). The statistics most commonly used are based on the site-frequency spectrum (Spence, Kamm, and Song 2016) and, as such, are nonphylogenetic in nature.

Here we show how to calculate the likelihood of multifurcating genealogies under Λ-coalescent models, given a function describing effective population size. We derive a new estimate of effective population size, which we call the multifurcating skyline plot, and extend this to longitudinal (serial) sampling. Using simulated data, we show that, in the case of $Beta(2-\alpha, \alpha)$ -coalescents, we can estimate the key α parameter that describes the propensity of the sample genealogy to contain multifurcations ($\alpha = 2$ corresponds to the Kingman coalescent, whilst $\alpha = 1$ represents the Bolthausen–Sznitman coalescent). Our analyses include data simulated under an empirically informed model of epidemiological super-spreading. We also show that effective population sizes can be estimated accurately. Note that this study aims to undertake inference from a single, pre-specified multifurcating genealogy. We leave for future work the problem of statistically inferring such multifurcating trees from an empirical gene sequence alignment.

2. Methods

2.1 Mathematical properties of Λ-coalescents

Coalescents are random trees that describe the genealogy of a small number of individuals sampled from a larger population. The class of Λ-coalescents is parametrized by a probability measure $\Lambda(dx)$ on the interval [0, 1]. In most cases, these probability measures will be absolutely continuous with respect to the Lebesgue measure, i.e. $\Lambda(dx) = \lambda(x)dx$, for some non-negative function $\lambda(x)$ such that $\int_0^1 \Lambda(dx) = \int_0^1 \lambda(x)dx = 1$.

Under the Kingman coalescent, sample trees are almost surely binary, However, trees under the Λ-coalescent are in general multifurcating (i.e. contain at least one node with degree larger than 3). The distribution of such trees can be described as follows: start with n sample lineages at time 0. Note that time is represented backwards hence time 0 represents the present (or, more generally, the time of the most recent tree tip). Consider, for each k-tuple of lineages with $2 \le k \le n$, an independent exponential random variable with parameter

$$\lambda_{n,k} = \int_0^1 x^{k-2}(1-x)^{n-k}\Lambda(dx). \tag{1}$$

Then, if T_n is the minimum of these random variables, T_n is exponentially distributed with parameter $\sum_{k=2}^{n} \binom{n}{k} \lambda_{n,k}$. At time T_n, choose the number K of coalescing lineages with probability

$$\mathbb{P}(K = k) = \frac{\binom{n}{k}\lambda_{n,k}}{\sum_{i=2}^{n}\binom{n}{i}\lambda_{n,i}}. \tag{2}$$

Finally, choose K lineages at random from the set of n extant lineages, and collapse them into a single lineage. Continue this process with the remaining $n - K + 1$ lineages until only one lineage is left. In other words, each k-tuple of lineages coalesces at rate $\lambda_{n,k}$, independently of all other tuples. To better understand the role played by the Λ measure in the distribution of multifurcations, Equation (2) can also be interpreted in the following way: assume that $\mu_{-2} = \int_0^1 x^{-2}\Lambda(dx) < \infty$, so that $F_1(dx) = x^{-2}\Lambda(dx)/\mu_{-2}$ is a probability measure. Then, at coalescence time T_n, let $p \in (0, 1]$ be sampled according to F_1; select coalescing lineages from the n extant lineages independently with probability P. For more details on this construction (sometimes dubbed the *paintbox* construction by analogy with the construction used by Kingman) see Pitman 1999.

In Equation (1), it is easy to see that when $\Lambda = \delta_0$, the Dirac mass at 0, then $\lambda_{n,k} = 0$ for $3 \le k \le n$, and $\lambda_{n,2} = 1$. Thus, if the probability measure Λ is concentrated solely at 0, then the Λ-coalescent process is strictly binary and identical to the Kingman coalescent. However, the paintbox construction scheme does not apply to this case, since $\int_0^1 \delta_0(dx)/x^2 = \infty$.

2.2 Variable population size

For a given set of n individuals sampled from a large populations, we can define the *coalescent effective population size*, N_e, which is the size of an ideal population exhibiting the same

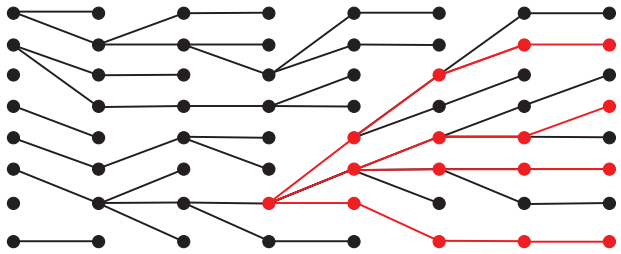

Figure 1. An ancestral genealogy (in red) of $n=4$ lineages sampled from a Cannings model with constant population size $N=8$ (black dots). The horizontal axis represents time and the lines indicate ancestry.

amount of genetic diversity under the Wright–Fisher reproduction model as the population under study. Indeed, one can recover a Kingman coalescent from the genealogy of n individuals in a Wright–Fisher model of size N by rescaling time by a factor of $N_e = N$ (Wakeley 2009). For more details on the different notions of effective population size see Sjödin et al. 2005.

For Λ-coalescents, the analog of the Wright–Fisher model is the *Cannings model* (Cannings 1974). In the Cannings model, all N individuals in one generation reproduce according to the same offspring distribution (see Fig. 1). If $\nu_1, \ldots, \nu_N$ are the number of offspring of individuals 1,..., N, respectively, we need to have $\nu_1 + \ldots + \nu_N = N$ in order to keep population size constant. A further condition is *exchangeability*, meaning that the distribution of the vector $(\nu_1, \ldots, \nu_N)$ is invariant under permutations. This implies in particular that all individuals have the same offspring distribution (i.e. the same propensity to reproduce). Hence, their common expectation is $\mathbb{E}[\nu_1] = \ldots = \mathbb{E}[\nu_N] = 1$. Let $\sigma^2(N)$ be their common variance. We can recover the Wright–Fisher model as a specific case of the Cannings model by taking $(\nu_1, \ldots, \nu_N)$ to be the multinomial distributions with parameters $(N;\ 1/N,\ \ldots, 1/N)$.

Suppose that the population size is large ($N \to \infty$). If the variance in offspring number converges to a fixed finite value as N increases ($\sigma^2(N) \to \sigma^2 < \infty$), then the genealogy of n lineages randomly sampled from the Cannings model of size N, rescaled by $N_e = N/\sigma^2$, still converges to the Kingman coalescent as in the Wright–Fisher case. If we relax this convergence condition, but still keep $\sigma^2(N) = o(N)$, under some additional technical conditions (see Theorem 3.1 in Sagitov 1999), then the sample genealogy converges to a Λ-coalescent when rescaled by the factor $N/\sigma^2(N)$. By analogy with the Wright–Fisher case, we can call this the Λ-effective population size. Since the Cannings model counts time in units of generations, this means that, when dealing with a real population that exhibits Cannings-like reproduction, it is necessary to count time in units of $N_e = N/\sigma^2(N)$ generations in order to observe Λ-coalescent-like behaviors. We can thus define a Λ-coalescent process in continuous time, given an effective population size function $(N_e(t),\ t \geq 0)$ by analogy with the variable population size coalescent described in (Griffiths and Tavare 1994).

Given a tree $\mathcal{T}$ with n tips we define $c(\mathcal{T})$ as the number of coalescences, that is, the number of internal nodes of $\mathcal{T}$. If $\mathcal{T}$ is a binary tree, then $c(\mathcal{T}) = n - 1$. We denote the coalescence times as $0 < t_1 < \ldots < t_{c(\mathcal{T})}$. For convenience, we denote present as $t_0 = 0$. During the interval $[t_{i-1}, t_i)$, with $1 \leq i \leq c(\mathcal{T})$, let n_i be the number of extant lineages. By definition, we always have $n_1 = n$. Finally, for each $1 \leq i \leq c(\mathcal{T})$, let k_i be the number of lineages involved in the coalescence at time t_i (see Fig. 2 for an example of this notation). Again by definition, we have $n_{i+1} = n_i - k_i + 1$.

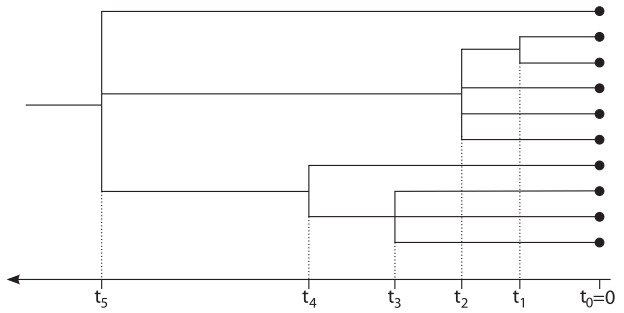

Figure 2. A multifurcating, ultrametric sample tree with $n=10$ tips. In this case, the number of lineages involved in each coalescence event is $k_1 = 2$, $k_2 = 4$, $k_3 = 3$, $k_4 = 2$, $k_5 = 3$. The number of extant lineages is $n_1 = 10$ during $[0, t_1)$, $n_2 = 9$ during $[t_1, t_2)$,..., $n_5 = 3$ during $[t_4, t_5)$.

The likelihood of a tree under the Λ-coalescent model, given the Λ distribution, is easy to write down thanks to the Markovian description of the process given above. It can in fact be decomposed in product form, since the waiting times between coalescent events are conditionally independent. Given the number of lineages n_i on an intercoalescent interval $[t_{i-1}, t_i)$, the waiting time is distributed as an exponential random variable with parameter

$$\sum_{j=2}^{n_i} \binom{n_i}{j} \lambda_{n_i,j}, \tag{3}$$

hence the likelihood of observing an interval of length $(t_i - t_{i-1})$ is

$$\sum_{j=2}^{n_i} \binom{n_i}{j} \lambda_{n_i,j} \exp\left(- \sum_{j=2}^{n_i} \binom{n_i}{j} \lambda_{n_i,j} (t_i - t_{i-1}) \right).$$

Then, at each coalescent time, the likelihood of seeing exactly k_i of the n_i lineages coalesce is, according to (2),

$$\mathbb{P}(K = k_i | \Lambda) = \frac{\binom{n_i}{k_i} \lambda_{n_i,k_i}}{\sum_{j=2}^{n_i} \binom{n_i}{j} \lambda_{n_i,j}}.$$

This gives, for the complete likelihood:

$$\mathbb{P}(\mathcal{T} | \Lambda) = \prod_{i=1}^{c(\mathcal{T})} \binom{n_i}{k_i} \lambda_{n_i,k_i} \times \exp\left(- \sum_{j=2}^{n_i} \binom{n_i}{j} \lambda_{n_i,j} (t_i - t_{i-1}) \right), \tag{4}$$

with the $\lambda_{n,k}$ depending on Λ as in (1). Note that when taking $\Lambda = \delta_0$ (i.e. the Kingman coalescent) the $\lambda_{n,k}$ are all zero except $\lambda_{2,2} = 1$, so that, as expected, we get a zero likelihood for nonbinary trees (i.e. trees with at least one $k_i > 2$) under the Kingman coalescent. Following Griffiths and Tavare (1994), we can now take into account the effective population size as a time-change of the Λ-coalescent, and write the likelihood of the tree given an effective population size function $(N_e(t),\ t \geq 0)$ under a Λ-coalescent model:

$$\mathbb{P}(\mathcal{T} | N_e, \Lambda) = \prod_{i=1}^{c(\mathcal{T})} \frac{\binom{n_i}{k_i} \lambda_{n_i,k_i}}{N_e(t_i)} \times \exp\left(- \int_{t_{i-1}}^{t_i} \frac{\sum_{j=2}^{n_i} \binom{n_i}{j} \lambda_{n_i,j}}{N_e(\tau)} d\tau \right). \tag{5}$$

In this formula, the population size function $(N_e(t),\ t \geq 0)$ is assumed to be deterministic (see Kaj and Krone 2003) for a treatment of Kingman coalescents with stochastically varying population size). The continuous-time model defined by (5) can be obtained as scaling limit of a Cannings model with variable population size, as shown in Theorem 2.2 of Möhle (2002). Indeed, this result, as in the constant population size case, shows that the effective population size function N_e is related to the parameters of the offspring distribution of a Cannings model through $N_e(t) = N(t)/\sigma^2(t)$, where $N(t)$ and $\sigma^2(t)$ are the census population size and the offspring variance, respectively, of the Cannings model at time t in the coalescent timescale. This relationship will be used later when considering the interpretation of estimated effective population sizes. Under the variable population size model, the change in timescale necessary to transform from generation counts to coalescent time is a little more involved than that for constant population size; details can be found in Möhle (2002) and Freund (2019).

3. Results

Using the likelihood formula (5) and given a time-scaled nonbinary tree $\mathcal{T}$, we will now show how to estimate certain features of the process that generated it using maximum likelihood. In Section 4, we will explore other uses of this likelihood, for instance its use as a tree prior in Bayesian phylogenetic inference frameworks, such as that implemented in BEAST (Drummond and Rambaut 2007).

3.1 Maximum-likelihood estimation of the Λ measure

We first consider the problem of estimating the Λ parameter directly from an observed genealogy. In the general case the relevant parameter space (i.e. the space of probability measures on $[0,1]$) is too large to enable direct inference of Λ (see Koskela, Jenkins, and Spanò 2018) for more advanced techniques). Therefore, we will limit ourselves here to the one-parameter family of $Beta(2-\alpha,\alpha)$-coalescents, with $\alpha \in (0,2)$. This corresponds to the case of Λ-coalescents where the Λ measure is the Beta distribution with parameters $2-\alpha$ and α, with $\alpha \in (0,2)$:

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

where $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ is the Beta function. By continuity, this can be extended to $\alpha = 2$, in which case the Beta distribution collapses to the Dirac measure at 0, thus recovering the Kingman coalescent. For $\alpha \in (0,2)$, the coalescence rates are given by:

$$\lambda_{n,k} = \frac{B(k-\alpha, n-k+\alpha)}{B(2-\alpha,\alpha)},\ 2 \leq k \leq n. \tag{6}$$

We show in Fig. 3 how the α parameter influences the coalescence probabilities, going from strictly binary coalescence events (for $\alpha = 2$) to having coalescence events involving a high number of lineages with high probability when $\alpha \to 0$. Indeed, the limit $\alpha = 0$ corresponds to the star-shaped coalescent, which has only a single coalescence event involving all lineages.

Beta-coalescent processes have been extensively studied for $\alpha \in [1,2]$, starting with Schweinsberg (2003), who described how they can be recovered from the genealogy of certain supercritical branching processes. Similarly, the connection with the so-called α-stable continuous-state branching processes (Birkner, Blath, and Eldon 2005) has enabled the study of many features of the family of Beta-coalescents (Berestycki, Berestycki, and Schweinsber 2007, 2008; Kersting, Schweinsberg, and Wakolbinger 2014). For $\alpha \in (0,1)$, the coalescent loses the so-called *coming down from infinity* property, which makes its mathematical analysis more difficult. For more details see e.g. Berestycki, Berestycki, and Limic 2010.

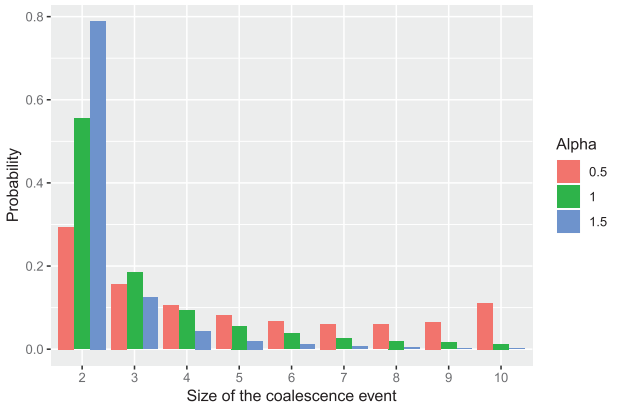

Figure 3. Probability of a coalescence event involving k lineages among $n = 10$ extant lineages in a $Beta(2-\alpha,\alpha)$-coalescent.

We investigated the problem of inferring the α parameter by using trees simulated under the $Beta(2-\alpha,\alpha)$-coalescent process with constant effective population size, for three different α values, specifically $\alpha = 1.2$; $\alpha = 1.5$; $\alpha = 1.8$. In each case, we simulated 1,000 trees with $n = 100$, 500, and 1,000 simultaneously sampled tips. We then estimated the α parameter independently for each tree by maximizing the likelihood (5):

$$\hat{\alpha}(\mathcal{T}) = \underset{\alpha\in[0,1]}{\operatorname{argmax}}\, \mathbb{P}(\mathcal{T}|N_e = 1, \Lambda = Beta(2-\alpha,\alpha)). \tag{7}$$

The results of this maximum-likelihood estimation are summarized in Table 1 and shown in Fig. 4. The procedure yields an effectively unbiased estimator, with decreasing variance as the number of tips increases. Furthermore, for a fixed number of tips, the lowest variance was obtained for true values of α close to 2, since the likelihood of trees with nonbinary nodes converges to 0 as $\alpha \to 2$, so that likelihood surfaces become more peaked. These results are encouraging because they indicate that features of the Λ measure might be identifiable from phylogenies.

3.2 Demographic inference using the skyline plot

We now turn to the problem of estimating ancestral effective population size, using a sample tree as data. Following the 'classic skyline plot' estimator (Pybus, Rambaut, and Harvey 2000), we assume that the duration of inter-coalescent intervals are known without error. We further assume that the degree of each internal node is known precisely. Although both of these variables are uncertain when trees are estimated from empirical data, we leave for future work the problem of incorporating uncertainty in node degree into phylodynamic inference (see Section 4).

It is possible, for a given probability measure Λ, and assuming that population size is constant between coalescent events, to maximize the likelihood function (5). This is made easy by the product form of the likelihood and yields the multifurcating skyline plot estimator:

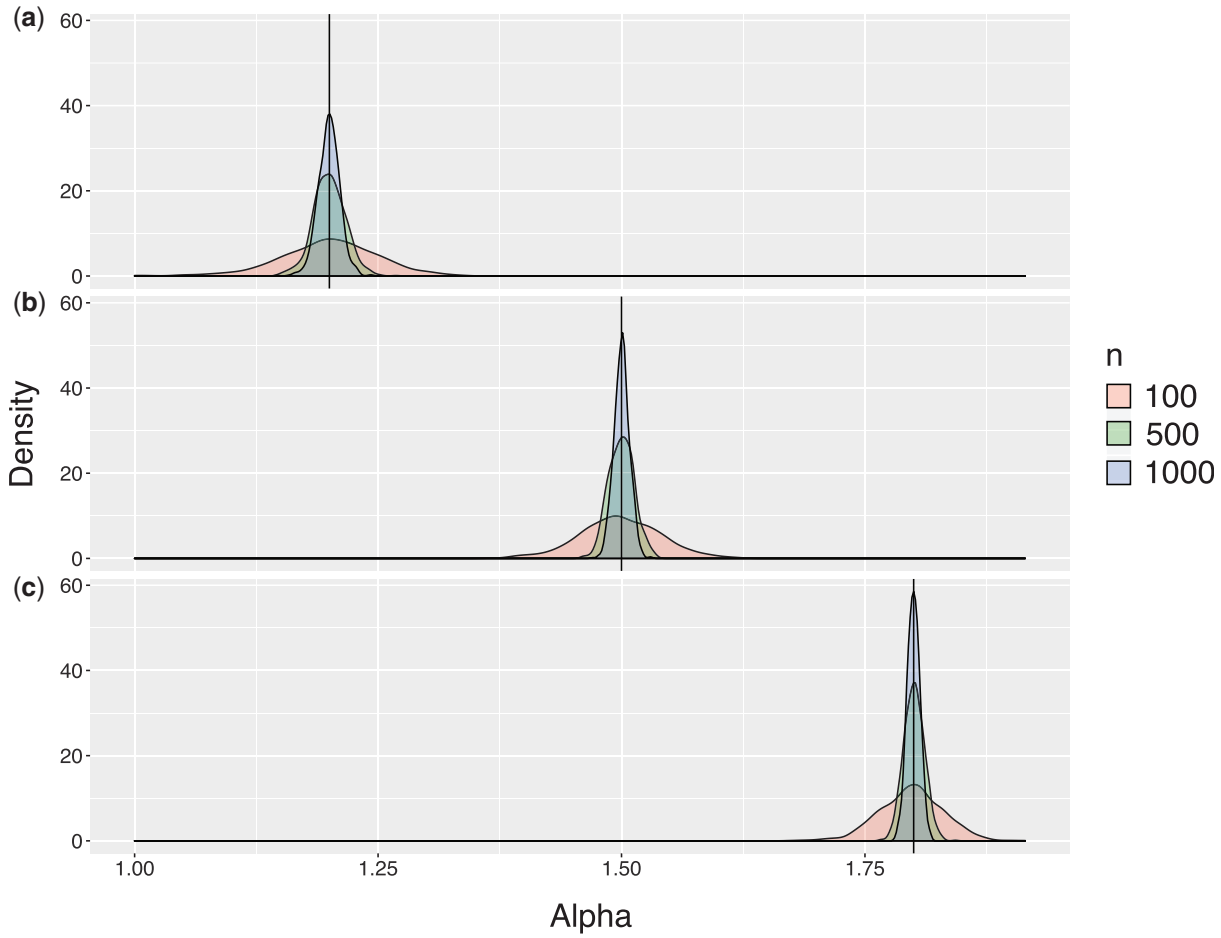

Figure 4. Empirical probability density of maximum-likelihood estimates of $\alpha \in [1,2]$ for 1,000 trees simulated under $Beta(2-\alpha,\alpha)$-coalescents, with (a) $\alpha = 1.2$, (b) $\alpha = 1.5$, and (c) $\alpha = 1.8$, respectively. The color of the density plots corresponds to the number n of tips in the simulated trees.

Table 1. Empirical mean, bias and variance of maximum-likelihood estimates of the α parameter of the Beta-coalescent process.

	Number of tips	Mean	Bias	Variance
α=1.2	100	1.2004	-4×10^{-4}	2.6×10^{-3}
	500	1.1999	-4.8×10^{-5}	2.8×10^{-4}
	1,000	1.1995	-4.7×10^{-4}	1.2×10^{-4}
$\alpha = 1.5$	100	1.499	-1.3×10^{-2}	1.7×10^{-3}
	500	1.499	-1.7×10^{-4}	1.7×10^{-4}
	1,000	1.500	1.7×10^{-4}	6.3×10^{-5}
$\alpha = 1.8$	100	1.797	-2.5×10^{-3}	1.1×10^{-3}
	500	1.800	4.2×10^{-4}	1.1×10^{-4}
	1,000	1.799	-2.0×10^{-4}	4.0×10^{-5}

$$\hat{N}_e(t) = \left(\sum_{j=2}^{n_i} \binom{n_i}{j} \lambda_{n_i,j}\right)(t_i - t_{i-1}),\ t \in [t_{i-1}, t_i). \tag{8}$$

Note that when $\Lambda = \delta_0$, the Dirac measure at 0, all the $\lambda_{n_i,j}$ are equal to 0, except for $j=2$, for which $\lambda_{n_i,2} = 1$, so that we recover the 'classic skyline plot' estimate (Pybus, Rambaut, and Harvey 2000) for binary trees:

$$\hat{N}_e(t) = \binom{n_i}{2}(t_i - t_{i-1}),\ t \in [t_{i-1}, t_i).$$

Finally, note that Equation (8) implies that the skyline plot estimator is expressed in the same units as the intercoalescent lengths $t_i - t_{i-1}$. Hence, if the underlying tree is scaled in real time units, it is necessary to divide the skyline plot estimate by the generation time to recover an estimate in population size units. For a detailed analysis of how varying generation time can affect skyline estimates see Volz 2012.

3.3 Composite internode intervals

The multifurcating skyline plot (8) is a piecewise constant function on inter-coalescent intervals and there can be a large number of such intervals when the number of tips is large, potentially leading to very noisy estimates. To mitigate this over-fitting, we can use the same interval-merging technique as that employed by the 'generalized skyline plot' (Strimmer and Pybus 2001). Given a threshold parameter $\epsilon > 0$, we consider all inter-coalescent intervals with length smaller than ϵ. We then join those intervals with their neighboring intervals earlier in time (closer to the root), starting with the interval closest to the tips. If the ensuing interval is still of length smaller than ϵ, we continue this procedure until there are no more intervals with length smaller than ϵ. Note that the degree of internal nodes is unchanged. This yields an interval partition of $[0, t_{c(\mathcal{T})})$ which we denote by $I_{1,\ell_1} \cup \ldots \cup I_{c_\epsilon(\mathcal{T}),\ell_{c_\epsilon(\mathcal{T})}}$. With this notation, $c_\epsilon(\mathcal{T}) \leq c(\mathcal{T})$ is the number of such composite intervals, and for $1 \leq i \leq c_\epsilon(\mathcal{T})$, ℓ_i is the number of inter-coalescent intervals joined together to form I_{i,ℓ_i}.

For each $1 \le i \le c(\mathcal{T})$, we then independently estimate a single effective population size value for the whole composite interval I_{i,ℓ_i}, using (5). Due to the product form of the likelihood, it is easy to see that the maximum-likelihood estimator for the composite interval I_{i,ℓ_i} is the mean of the maximum-likelihood estimators (8) for each of the ℓ_i inter-coalescent intervals (note that the original 'generalized skyline plot' of Strimmer and Pybus (2001) uses a method-of-moments estimator, which leads to a harmonic mean rather than an arithmetic mean).

Thus this composite-interval approach can compute an estimate of effective population size change with fewer parameters. As suggested in Strimmer and Pybus (2001), we can then optimize over $\epsilon \ge 0$ by a model selection statistic such as the corrected Akaike Information Criterion (Hurvich and Tsai 1989):

$$\mathrm{AICc}(\epsilon) = 2c_\epsilon(\mathcal{T}) - \log L_\epsilon + \frac{2c_\epsilon(\mathcal{T})(c_\epsilon(\mathcal{T})+1)}{n - c_\epsilon(\mathcal{T}) - 1} \tag{9}$$

where L_ϵ is the likelihood of the tree given the composite interval skyline estimate with parameter ϵ, as computed by (5).

3.4 Serially sampled trees

In some biological contexts, such as rapidly evolving pathogens or ancient DNA, sequences are sampled at different times and measurable amounts of sequence change occur between the sampling times (Drummond et al. 2003). As described in Rodrigo et al. (1999), coalescent estimates can be extended to this framework, by making the assumption that the sampling process is independent of the population dynamics. For the purpose of simplicity, in this paper we also make this independence assumption. We further assume that effective population size does not change at sampling times and that sampling never occurs at coalescence times.

To be precise, let us consider an inter-coalescent interval $[t_{i-1}, t_i)$, and assume we have n_i extant lineages at time t_{i-1}. Assume we have sampling times $t_{i-1} = s_0^i < s_1^i < \ldots < s_k^i < t_i$, and for $1 \le j \le k$, let $n(j)$ be the number of extant lineages on the interval ending with s_j^i (with $n(1) = n_i$). Finally, let $n(k+1)$ be the number of lineages on the coalescence interval $[s_k, t_i)$. Figure 5 provides a graphical example of a serially sampled tree and its associated indices.

When computing the likelihood of a serially sampled tree under a specified Λ-coalescent model, given an effective population size function, we need to take into account the intervals of the form $[t_{i-1}, s_1^i)$ and $[s_j^i, s_{j+1}^i)$ that end with a sampling event. On those intervals there are no coalescences, hence we need to consider the likelihood of no coalescences occurring during that time. The contribution to the likelihood of the inter-coalescent interval $[t_{i-1}, t_i)$ is then:

$$\frac{\binom{n(k+1)}{k_i}\lambda_{n(k+1),k_i}}{N_e} \times \exp\left(-\frac{\sum_{p=2}^{n(k+1)}\binom{n(k+1)}{p}\lambda_{n(k+1),p}\left(t_i - s_k^i\right)}{N_e}\right) \times \prod_{j=1}^{k}\exp\left(-\frac{\sum_{p=2}^{n(j)}\binom{n(j)}{p}\lambda_{n(j),p}\left(s_j^i - s_{j-1}^i\right)}{N_e}\right) \tag{10}$$

Again, maximizing this in N_e gives the maximum-likelihood skyline estimator on the interval $[t_{i-1}, t_i)$:

$$\hat{N}_e(t) = \sum_{p=2}^{n(k+1)}\binom{n(k+1)}{p}\lambda_{n(k+1),p}\left(t_i - s_k^i\right) + \sum_{j=1}^{k}\sum_{p=2}^{n(j)}\binom{n(j)}{p}\lambda_{n(j),p}\left(s_j^i - s_{j-1}^i\right), \quad t \in [t_{i-1}, t_i). \tag{11}$$

3.5 Inference using the multifurcating skyline plot

We evaluated the ability of the multifurcating skyline plot to estimate N_e by applying it to trees simulated under the Beta-coalescent process. We considered two different demographic histories (constant population size and exponential growth) and two values of the interval length threshold parameter ϵ. We used the same α parameter in both cases, namely $\alpha = 1.5$. The results are shown in Fig. 6.

The results are in line with well-known properties of skyline estimators (Strimmer and Pybus 2001; Drummond et al. 2005): they recover fluctuations in effective population size reasonably well, at the cost of being quite noisy. As expected, noise decreases as the interval-merging parameter ϵ increases (Fig. 6).

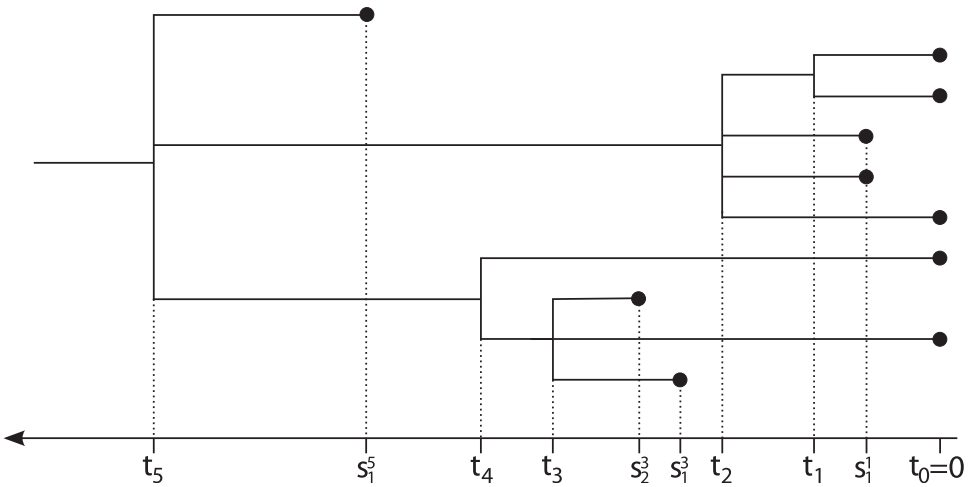

Figure 5. A serially sampled multifurcating tree, with 10 tips in total. There are $c(\mathcal{T}) = 5$ coalescent events. During the inter-coalescent interval $[t_2, t_3)$, the number of extant lineages is (going back in time) $n_2 = n(1) = 3, n(2) = 4$, and $n(3) = 5$. Using the composite-interval notation, assuming for example that $t_4 - t_3 < \epsilon < t_2 - t_1$, the ensuing interval partition would have a I_{4,ℓ_4} component, with $\ell_4 = 2$ corresponding to the two intervals $[t_3, t_4)$ and $[t_4, t_5)$ having been joined together. There would be $c_\epsilon(\mathcal{T}) = 4$ intervals in the merged interval partition.

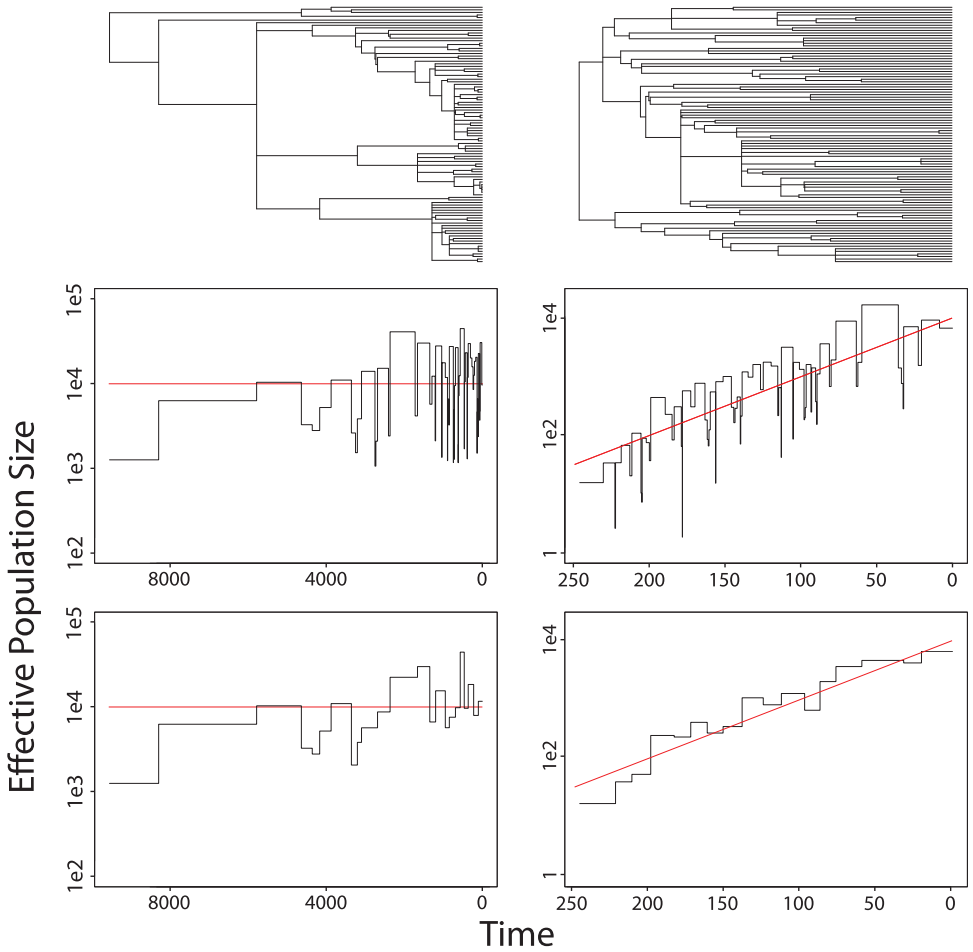

Figure 6. Top row: individual trees simulated with 100 tips under a *Beta*$(0.5, 1.5)$-coalescent, under constant (left) and exponential population size (right). Red lines represent the true effective population size function. Middle row: multifurcating skyline plots of the trees, with fixed $\alpha = 1.5$ and with $\epsilon = 0$. Bottom row: generalized multifurcating skyline plots, with threshold length parameter $\epsilon = 100$ (left) and $\epsilon = 10$ (right). Time is expressed in number of generations, and the estimated effective population size is in population size units.

3.6 Estimation of superspreading parameters

To demonstrate the potential usefulness of the multifurcating skyline framework, we generated phylogenetic trees using a previously published individual-based simulation model (Li, Grassly, and Fraser 2017) that implements a well-known model of epidemiological superspreading (Lloyd-Smith et al. 2005). These simulated trees were then analyzed using the Λ-coalescent model presented above.

The Lloyd-Smith superspreading model assumes that each infectious individual i gives rise to a Poisson-distributed number Z_i of secondary infections, with random parameter ν_i. These ν_i are in turn distributed according to a Gamma distribution with shape R_0/k and scale k, where $R_0 > 0$ is the population-level basic reproductive number, and $k > 0$ is a parameter characterizing the degree of superspreading prevalent in the epidemic. This gives rise to the following (negative binomial) individual transmission distribution:

$$\mathbb{P}(Z = n) = \binom{n+k-1}{n}\left(1 - \frac{R_0}{R_0+k}\right)^k \left(\frac{R_0}{R_0+k}\right)^n, \qquad (12)$$

for $n \geq 0$. As a consequence, each infectious individual transmits the infection to an average of $\mathbb{E}[Z] = R_0$ individuals, with a variance $\mathrm{Var}(Z) = R_0(1 + R_0/k)$. Hence, for a given value of R_0, low values of k correspond to highly variable onward transmission, such that some individuals (the so-called *superspreaders*) are responsible for a disproportionate number of transmission events. The Lloyd-Smith model has been fitted to observed secondary cases obtained from real epidemics using contact tracing. Estimated parameter values range from $(R_0 = 1.63,\ k = 0.16)$ for the 2003 SARS outbreak in Singapore to $(R_0 = 1.5,\ k = 5.1)$ for an Ebola outbreak in Uganda. Generally, k values lower than 1 are considered to be evidence of high levels of superspreading.

In order to explore the effect of superspreading dynamics on multifurcating sample phylogenies, we simulated trees under the Lloyd-Smith model using the EpiGenR R package[1] (Li, Grassly, and Fraser 2017). This package tracks the transmission history of an infected population within a host population of constant size N. Each infectious individual remains infectious for an exponentially distributed time, then gives rise to a number of secondary cases distributed according to (12). Details of the implementation can be found in the Supplementary Data section of Li, Grassly, and Fraser (2017).

We investigated three superspreading scenarios, namely $k = 0.15$, $k = 0.37$, and $k = 5$, which represent high, moderate, and low levels of superspreading, respectively. These values are consistent with the published parameter estimates for SARS

and Ebola outbreaks (Lloyd-Smith et al. 2005; Lau et al. 2017). We chose the same value of R_0 =2.5 for all three scenarios. In each scenario, we simulated 100 epidemics in a host population of size $N = 10{,}000$. A fraction s of the tips in the resulting simulated complete transmission trees were then subsampled randomly to generate sample phylogenies. This sampling was undertaken uniformly across all tips of the tree, so that most tips were sampled during the times where prevalence is high. We explored four values of sampling intensity; $s = 0.01$, 0.05, 0.1, and 0.5.

We then estimated the $\alpha \in (0, 2)$ parameter of the $Beta(2-\alpha, \alpha)$ -coalescent using maximum-likelihood estimation, accounting for demographic variation by using the multifurcating skyline plot as introduced above.

The results of this simulation analysis are shown in Fig. 7. In all sampling regimes, we find a positive relationship between the superspreading parameter k and the α parameter of the *Beta*-coalescent. Even for low sampling ($s = 0.01$, corresponding to trees with $\leq$100 tips), estimated α was significantly higher when $k = 5$ than when $k = 0.15$ and 0.37. This shows that the timing of coalescent events in multifurcating Λ-coalescent trees contains information about the extent of superspreading in an infectious population. This is consistent with the findings of Li, Grassly, and Fraser (2017), who used particle filtering techniques to show that there is information about transmission heterogeneity in estimated phylogenies that is not present in the time series data alone. Furthermore, Fig. 7 shows the multifurcating phylogenetic models used here perform adequately under high levels of sampling. Indeed, by sampling a large portion of the infectious population, there is an increased probability of seeing deep and large multifurcations in the phylogeny, which likely drives down estimates of α.

4. Discussion

We have shown that Λ-coalescents can be used to infer demographic trends from multifurcating trees, extending the range of 'skyline plot'-like estimators beyond binary trees and the standard Kingman coalescent. Using simulated data, we showed that the α parameter of the $Beta(2-\alpha, \alpha)$ family of Λ-coalescents can be accurately estimated, which interpolates between the Kingman coalescent ($\alpha = 2$) and the Bolthausen–Sznitman coalescent ($\alpha = 1$). We introduce the multifurcating skyline plot, which estimates effective population size from time-scaled nonbinary trees, the tips of which may be sampled either longitudinally or concurrently. We validated this estimation method on simulated trees.

We applied the framework of Λ-coalescents to trees simulated using two different approaches, including an individual-based simulation of epidemic superspreading. Crucially, we demonstrated that the temporal distribution of bifurcations and multifurcations in a sample tree contains information about the extent of superspreading. However, the simulation studies presented here neglect the role played by phylogenetic uncertainty, as we assumed perfect knowledge of the transmission trees. Further work is required to implement the inference of multifurcating trees from sequence alignments in Bayesian phylogenetic frameworks such as BEAST. This will necessitate the definition of tree operators capable of exploring the larger space of nonbinary trees; whether this can be done without affecting computational performance remains an open question. Joint evaluation of molecular clock phylogenetic likelihoods and multifurcating tree prior probabilities has the potential to discriminate between genuine multifurcations, and short tree branches on which no mutations are observed.

Several popular Bayesian implementations of the skyline plot approach exist, which treat the skyline plot likelihood as a 'prior distribution' on trees. These methods include the 'Bayesian skyline plot' (Drummond et al. 2005), which uses the composite-interval procedure described in this paper to reduce noise, and the skyride (Minin, Bloomquist, and Suchard 2008) and skygrid (Gill et al. 2013) plots, which use sophisticated, time-aware, smoothing procedures to penalize population size changes. We expect similar approaches could be applied to the multifurcating skyline plot. Recent work (Möller, du Plessis, and Stadler 2018) has illustrated that mis-specification of the tree prior during

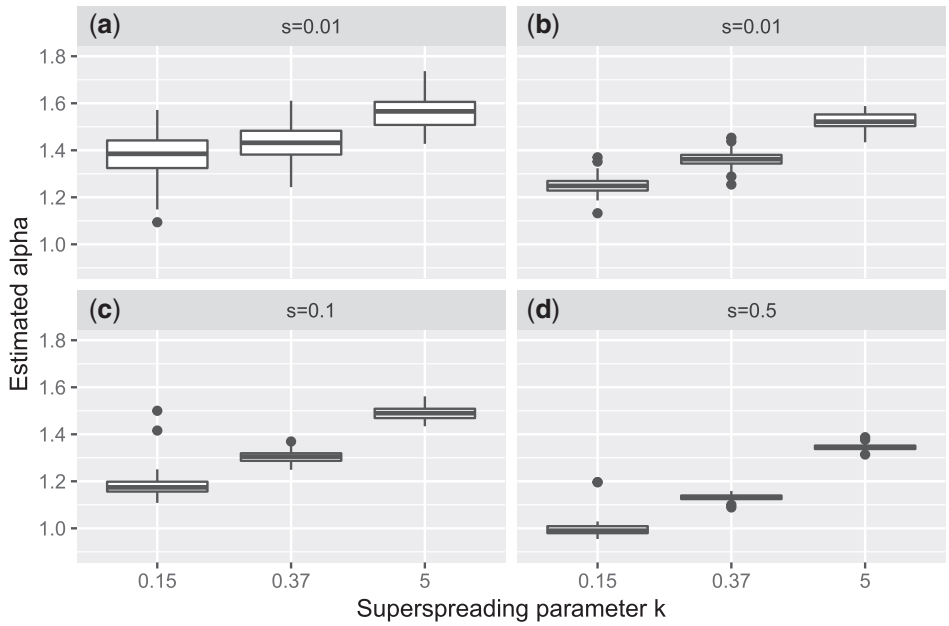

Figure 7. Estimated $\alpha \in (0, 2)$ for phylogenies simulated under the Lloyd-Smith superspreading model, for R_0=2.5 and $k = 0.15$, 0.37, and 5, respectively. Increasing levels of tip sampling are shown: (a) $s = 0.01$, (b) $s = 0.05$, (c) $s = 0.1$ and (d) $s = 0.5$.

Bayesian phylogenetic inference may lead to erroneous estimation of other parameters, highlighting the need to implement a wider range of computationally tractable tree priors. The multifurcating skyline plot may help to address this deficit. We hope it will prove useful in the analysis of populations that exhibit patterns of propagation leading to true polytomies in their phylogeny, such as superspreading (as presented here), strong selective regimes (Neher and Hallatschek 2013), oversampling (Bhaskar, Clark, and Song 2014) or even adaptive radiation (Schluter 2000). The estimated parameters of the Λ distribution might be informative about these propagation processes, which are not taken into account in current phylodynamic approaches. However, we accept the possibility that the *Beta*-coalescents used here are not the best models for multifurcating phylodynamic inference and that other Λ-coalescent families can be potentially used to estimate parameters of epidemiological interest.

Although skyline plot methods based on the Kingman coalescent are commonplace in phylodynamic analysis, the actual values of estimated effective population size are not often interpreted directly. In some cases, 'true', census population size and effective population size are related only through nonlinear relations. For example, it was shown (Frost and Volz 2010; Volz 2012) that under SIR-type epidemiological models (Kingman) coalescent effective population size through time $N_e(t)$ evolves proportionally to $I_t^2/f_{SI}(t)$, where $f_{SI}(t)$ is the rate at which susceptible individuals become infected (typically, $f_{SI}(t) \propto S_t I_t/N$). In this case, the relationship between effective population size $N_e(t)$ and disease incidence I_t at time t is different at the beginning and end of the epidemic. The multifurcating skyline methods will pose similar problems for interpretation: the actual value of the estimate is affected by the Λ distribution in a nonlinear (and not completely understood) way. It is thus necessary to caution against improper interpretations of the Λ-effective population size, as defined in this paper and as computed by the multifurcating skyline plot. Other authors (Hall, Woolhouse, and Rambaut 2016) have already pointed out the dangers of thinking about effective population size as a number of individuals and have advised to use it as a measure of genetic diversity of the population instead. Such advice should also be heeded in the context of multifurcating coalescents, for which the relationship between census population size and effective population size is even more complicated.

Acknowledgements

The authors wish to thank Tim Vaughan and Christophe Fraser for useful conversations that led to the overall improvement of this manuscript.

Data availability

The scripts used in this paper for computing multifurcating skyline plots and maximum-likelihood estimations of the α parameter were published as an R package available at https://github.com/phoscheit/LambdaSkyline. We also included the EpiGenR scripts used for the simulation of phylogenies under the Lloyd-Smith superspreading model.

Funding

This work was supported by the European Research Council under the European Commission Seventh Framework Programme (FP7/2007-2013)/European Research Council grant agreement 614725-PATHPHYLODYN. P.H. has received the support of the EU in the framework of the Marie-Curie FP7 COFUND People Programme, through the award of an AgreenSkills fellowship under grant agreement no. 267196.

Conflict of interest: None declared.

References

Berestycki, J., Berestycki, N., and Limic, V. (2010) 'The Λ-Coalescent Speed of Coming Down From Infinity', *The Annals of Probability*, 38: 207–33.

——, ——, and Schweinsberg, J. (2007) 'Beta-Coalescents and Continuous Stable Random Trees', *The Annals of Probability*, 35: 1835–87.

——, ——, and —— (2008) 'Small-Time Behavior of Beta Coalescents', *Annales de L'Institut Henri Poincare (B) Probability and Statistics*, 44: 214–38.

Bhaskar, A., Clark, A. G., and Song, Y. S. (2014) 'Distortion of Genealogical Properties When the Sample Is Very Large', *Proceedings of the National Academy of Sciences*, 111: 2385–90.

Biek, R. et al. (2015) 'Measurably Evolving Pathogens in the Genomic Era', *Trends in Ecology & Evolution*, 30: 1–8.

Birkner, M. et al. (2005) 'Alpha-Stable Branching and Beta-Coalescents', *Electronic Journal of Probability*, 10: 303–25.

——, Blath, J., and Eldon, B. (2013) 'Statistical Properties of the Site-Frequency Spectrum Associated with Lambda-Coalescents', *Genetics*, 195: 1037–53.

Cannings, C. (1974) 'The Latent Roots of Certain Markov Chains Arising in Genetics: A New Approach, I. Haploid Models', *Advances in Applied Probability*, 6: 260.

Drummond, A. J. et al. (2003) 'Measurably Evolving Populations', *Trends in Ecology & Evolution*, 18: 481–8.

——, and Rambaut, A. (2007) 'BEAST: Bayesian Evolutionary Analysis by Sampling Trees', *BMC Evolutionary Biology*, 7: 214–8.

—— et al. (2005) 'Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences', *Molecular Biology and Evolution*, 22: 1185–92.

Durrett, R., and Schweinsberg, J. (2004) 'Approximating Selective Sweeps', *Theoretical Population Biology*, 66: 129–38.

Eldon, B. et al. (2015) 'Can the Site-Frequency Spectrum Distinguish Exponential Population Growth from Multiple-Merger Coalescents?' *Genetics*, 199: 841.

Freund, F. (2019) 'Cannings Models, *Populations Size Changes and Multiple-Merger Coalescents*', 1–18. arXiv 1902.02155.

Frost, S. D. W., and Volz, E. M. (2010) 'Viral Phylodynamics and the Search for an 'Effective Number of Infections'', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365: 1879–90.

Gill, M. S. et al. (2013) 'Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci', *Molecular Biology and Evolution*, 30: 713–24.

Griffiths, R. C., and Tavare, S. (1994) 'Sampling Theory for Neutral Alleles in a Varying Environment', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 344: 403–10.

Hall, M. D., Woolhouse, M. E., and Rambaut, A. (2016) 'The Effects of Sampling Strategy on the Quality of Reconstruction of Viral Population Dynamics Using Bayesian Skyline Family Coalescent Methods: A Simulation Study', *Virus Evolution*, 2: 1–14.

Hallatschek, O. (2018) 'Selection-Like Biases Emerge in Population Models With Recurrent Jackpot Events', *Genetics*, 210: 1053–73.

Ho, S. Y. W., and Shapiro, B. (2011) 'Skyline-Plot Methods for Estimating Demographic History from Nucleotide Sequences', *Molecular Ecology Resources*, 11: 423–34.

Hudson, R. R., and Kaplan, N. L. (1988) 'The Coalescent Process in Models with Selection and Recombination', *Genetical Research*, 120: 831–40.

Hurvich, C. M., and Tsai, C. L. (1989) 'Regression and Time Series Model Selection in Small Samples', *Biometrika*, 76: 297–307.

Kaj, I., and Krone, S. M. (2003) 'The Coalescent Process in a Population With Stochastically Varying Size', *Journal of Applied Probability*, 40: 33–48.

Kaplan, N. L., Darden, T., and Hudson, R. R. (1988) 'The Coalescent Process in Models With Selection', *Genetics*, 120: 819–29.

Kersting, G., Schweinsberg, J., and Wakolbinger, A. (2014) 'The Evolving Beta Coalescent', *Electronic Journal of Probability*, 19: 1–28.

Kingman, J. F. C. (1982) 'The Coalescent', *Stochastic Processes and Their Applications*, 13: 235–48.

Koskela, J. (2018) 'Multi-Locus Data Distinguishes between Population Growth and Multiple Merger Coalescents', *Statistical Applications in Genetics and Molecular Biology*, 17:

——, Jenkins, P. A., and Spanò, D. (2018) 'Bayesian Non-Parametric Inference for Lambda-Coalescents: Posterior Consistency and a Parametric Method', *Bernoulli*, 24: 2122–53.

Lau, M. S. Y. et al. (2017) 'Spatial and Temporal Dynamics of Superspreading Events in the 2014–2015 West Africa Ebola Epidemic', *Proceedings of the National Academy of Sciences*, 114: 2337–42.

Li, L. M., Grassly, N. C., and Fraser, C. (2017) 'Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series', *Molecular Biology and Evolution*, 34: 2982–95.

Lloyd-Smith, J. O. et al. (2005) 'Superspreading and the Effect of Individual Variation on Disease Emergence', *Nature*, 438: 355–9.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008) 'Smooth Skyride Through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics', *Molecular Biology and Evolution*, 25: 1459–71.

Möhle, M. (2002) 'The Coalescent in Population Models With Time-Inhomogeneous Environment', *Stochastic Processes and Their Applications*, 97: 199–227.

Möller, S., du Plessis, L., and Stadler, T. (2018) 'Impact of the Tree Prior on Estimating Clock Rates During Epidemic Outbreaks', *Proceedings of the National Academy of Sciences*, 115: 4200.

Neher, R. A., and Hallatschek, O. (2013) 'Genealogies of Rapidly Adapting Populations', *Proceedings of the National Academy of Sciences of the United States of America*, 110: 437–42.

Pitman, J. (1999) 'Coalescents With Multiple Collisions', *The Annals of Probability*, 27: 1870–902.

Pybus, O. G., and Rambaut, A. (2009) 'Evolutionary Analysis of the Dynamics of Viral Infectious Disease', *Nature Reviews Genetics*, 10: 540–50.

——, ——, and Harvey, P. H. (2000) 'An Integrated Framework for the Inference of Viral Population History from Reconstructed Genealogies', *Genetics*, 155: 1429–37.

Rodrigo, A. G. et al. (1999) 'Coalescent Estimates of HIV-1 Generation Time in Vivo', *Proceedings of the National Academy of Sciences of the United States of America*, 96: 2187–91.

Sackman, A. M., Harris, R. B., and Jensen, J. D. (2019) 'Inferring Demography and Selection in Organisms Characterized by Skewed Offspring Distributions', *Genetics*, 211: 1019–28.

Sagitov, S. (1999) 'The General Coalescent with Asynchronous Mergers of Ancestral Lines', *Journal of Applied Probability*, 36: 1116–25.

Sargsyan, O., and Wakeley, J. (2008) 'A Coalescent Process with Simultaneous Multiple Mergers for Approximating the Gene Genealogies of Many Marine Organisms', *Theoretical Population Biology*, 74: 104–14.

Schluter, D. (2000) *The Ecology of Adaptive Radiation*. Oxford: Oxford University Press.

Schweinsberg, J. (2003) 'Coalescent Processes Obtained from Supercritical Galton–Watson Processes', *Stochastic Processes and Their Applications*, 106: 107–39.

Sjödin, P. et al. (2005) 'On the Meaning and Existence of an Effective Population Size', *Genetics*, 169: 1061–70.

Spence, J. P., Kamm, J. A., and Song, Y. S. (2016) 'The Site Frequency Spectrum for General Coalescents', *Genetics*, 202: 1549–61.

Strimmer, K., and Pybus, O. G. (2001) 'Exploring the Demographic History of DNA Sequences Using the Generalized Skyline Plot', *Molecular Biology and Evolution*, 18: 2298–305.

Takahata, N. (1988) 'The Coalescent in Two Partially Isolated Diffusion Populations', *Genetical Research*, 52: 213.

Volz, E. M. (2012) 'Complex Population Dynamics and the Coalescent under Neutrality', *Genetics*, 190: 187–201.

Wakeley, J. (2009) *Coalescent Theory: An Introduction*. Colorado: Roberts & Company Publishers Greenwood Village.