Journal of Theoretical Biology 548 (2022) 111186

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

The bounded coalescent model: Conditioning a genealogy on a minimum root date

Jake Carson[a], Alice Ledda[b], Luca Ferretti[c], Matt Keeling[a], Xavier Didelot[d,*]

[a] Mathematics Institute, University of Warwick, United Kingdom
[b] HCAI, Fungal, AMR, AMU & Sepsis Division, UK Health Security Agency, United Kingdom
[c] Big Data Institute, University of Oxford, United Kingdom
[d] Department of Statistics and School of Life Sciences, University of Warwick, United Kingdom

ARTICLE INFO

Article history:
Received 4 January 2022
Revised 5 May 2022
Accepted 2 June 2022
Available online 10 June 2022

Keyword:
Coalescent model
Most recent common ancestor
Heterochronous sampling
Phylogenetics
Phylodynamics

ABSTRACT

The coalescent model represents how individuals sampled from a population may have originated from a last common ancestor. The bounded coalescent model is obtained by conditioning the coalescent model such that the last common ancestor must have existed after a certain date. This conditioned model arises in a variety of applications, such as speciation, horizontal gene transfer or transmission analysis, and yet the bounded coalescent model has not been previously analysed in detail. Here we describe a new algorithm to simulate from this model directly, without resorting to rejection sampling. We show that this direct simulation algorithm is more computationally efficient than the rejection sampling approach. We also show how to calculate the probability of the last common ancestor occurring after a given date, which is required to compute the probability density of realisations under the bounded coalescent model. Our results are applicable in both the isochronous (when all samples have the same date) and heterochronous (where samples can have different dates) settings. We explore the effect of setting a bound on the date of the last common ancestor, and show that it affects a number of properties of the resulting phylogenies. All our methods are implemented in a new R package called BoundedCoalescent which is freely available online.

1. Introduction

The coalescent model is a stochastic process that describes the ancestry of a sample of individuals within a population (Kingman, 1982; Kingman, 1982). Conditioning the most recent common ancestor of the sample to be after a certain date results in a model called the "bounded coalescent" model and which was first mentioned in a model unifying gene duplication, loss and coalescence (Rasmussen and Kellis, 2012). In this context, the bound condition is used to deal with incomplete lineage sorting, which could cause a gene tree to be incongruent with the species tree (Maddison, 1997; Maddison and Knowles, 2006). Consequently, the bounded coalescent model is used in many multi-species coalescent models, to enforce the full coalescence of members of a same species before the speciation event (Mallo et al., 2016; Du et al., 2019; Hill et al., 2020; Li et al., 2021). The bounded coalescent model is also used in work on homologous recombination resulting in ancestral recombination graphs (Ferretti et al., 2013; Rasmussen et al., 2014). Furthermore, the bounded coalescent model is useful to perform pathogen transmission analysis from genetic data whilst accounting for within–host diversity (Didelot et al., 2014; Didelot et al., 2017). In this case, setting a bound on the coalescent process equates to assuming a complete transmission bottleneck, so that all pathogen lineages within an infected individual need to coalesce before the host became infected. This results in a simpler relationship between the transmission tree of who-infected-whom and the genealogy of the pathogen sampled from the infected individuals. Despite this increasingly frequent use of the bounded coalescent model in several different biological research fields, the consequences of imposing a minimum on the date of the last common ancestor have not been formally investigated. In particular, under the standard coalescent model the waiting times between coalescent events are independent whereas this is no longer the case in the bounded coalescent model (Rasmussen and Kellis, 2012) since an increase in a waiting time needs to be compensated by a decrease in other waiting times to satisfy the bound condition.

In this paper, we start by defining the bounded coalescent model formally in both isochronous and heterochronous sampling settings. In the former all individuals are sampled at the same time,

[*] Corresponding author.

https://doi.org/10.1016/j.jtbi.2022.111186
0022-5193/© 2022 The Author(s). Published by Elsevier Ltd.

so that the genealogy is ultrametric (all leaves have the same distance to the root), whereas in the latter the individuals are sampled at different times, so that the genealogy is not ultrametric (Rambaut, 2000). We show how the probability density of any genealogy can be computed under the bounded coalescent model with a given effective population size and bound time. This requires to first compute the probability of having the bound property occurring under a standard unbounded coalescent model, and we show how this can be computed efficiently. We also present a new algorithm for the simulation of genealogies under the bounded coalescent model given the sample number and dates, the effective population size and the bound time. Our algorithm can simulate directly from the bounded coalescent model, unlike previously described approaches which used rejection sampling on trees simulated from the standard unbounded coalescent model (Didelot et al., 2014; Mallo et al., 2016). These new algorithms to calculate the probability density of a genealogy and to simulate directly under the bounded coalescent model are both useful to perform inference under the model. Finally, we investigate a number of properties of the genealogies arising from the bounded coalescent model and how they differ from the unbounded coalescent model.

2. Definitions and notations

Coalescent models are typically derived from forward-in-time population models, such as the Wright-Fisher model (Wright, 1931; Fisher, 1930), the Moran model (Moran, 1958) and the Cannings exchangeable model (Cannings, 1974). The Wright-Fisher model assumes that a population evolves through non-overlapping generations, with each individual in each generation having a random uniformly distributed ancestor in the previous generation, independent of the others. Under a constant population size N, the probability that two individuals have the same ancestor in the previous generation is $1/N$, so that the number of generations back-in-time until a common ancestor is found follows a Geometric distribution with mean N. This can be converted to real time by multiplying the number of generations by the generation interval T_g, resulting in the effective population size $N_e = NT_g$. Where N_e is sufficiently large, we can instead use the Kingman coalescent model (Kingman, 1982; Kingman, 1982), which replaces the Geometric distribution with an Exponential distribution and provides a continuous-time equivalent.

Under the standard coalescent model, the waiting time Δt for a coalescence with a lineages has for probability density

$$f(\Delta t|a) = \frac{a(a-1)}{2N_e} \exp\left(-\frac{a(a-1)}{2N_e}\Delta t\right), \tag{1}$$

since each pair of lineages coalesces at rate $1/N_e$ and there are $\frac{a(a-1)}{2}$ pairs of lineages.

In the most commonly used coalescent setting, L samples are taken simultaneously (isochronous) so that their ancestral process is simply made of the $L-1$ coalescent events occurring back in time until the most recent common ancestor (MRCA) of the L samples is found. The probability density of the coalescent dates in this isochronous setting can therefore be computed as the product of $L-1$ terms given by Eq. (1).

Here we consider a frequently used extension of this setting, in which the leaves are taken at different dates (heterochronous) $t_1 < t_2 < \ldots < t_K$ (Drummond et al., 2002; Drummond et al., 2003). We define L_k as the number of leaves taken at date $t_k, L = \sum_{k=1}^{K} L_k$ as the total number of leaves, and $\tau_1 < \tau_2 < \ldots < \tau_{L-1}$ as the coalescent node times. Here and throughout this manuscript time is measured in the forward direction, so that for example τ_1 is the date of the MRCA, and t_K is the date of the most recent sample. Furthermore we define the number of extant lineages at time t as

$$A(t) = \sum_{i=1}^{K} \mathbb{I}(t_i \geqslant t)L_i - \sum_{j=1}^{L-1} \mathbb{I}(\tau_j > t), \tag{2}$$

so that if t is a coalescence time, $A(t)$ is the number of lineages that could have coalesced. These definitions are illustrated using an example phylogeny in Fig. 1. Note that the isochronous case is a special case of the heterochronous case in which $K = 1$ and $L_1 = L$.

Letting $s_1, \ldots, s_{K+L-1}$ be the ordered union of the leaves and coalescent node times, and $D_k = \{t_k, L_k\}$ denote the combined sampling information, the probability density of the ancestor dates in a heterochronous setting can be calculated as follows (Drummond et al., 2002):

$$f(\tau_{1:L-1}|D_{1:K}) = \left(\prod_{j=1}^{L-1} \frac{A(\tau_j)(A(\tau_j)-1)}{2N_e}\right) \cdot \left(\prod_{i=1}^{K+L-2} \exp\left(-\frac{A(s_i)(A(s_i)-1)}{2N_e}(s_{i+1}-s_i)\right)\right), \tag{3}$$

where we use the notations $\tau_{1:L-1} = \tau_1, \ldots, \tau_{L-1}$ and $D_{1:K} = D_1, \ldots, D_K$.

In the bounded coalescent model we have the additional requirement that the lineages must coalesce to their MRCA before some specified bound time t^*, so that $\tau_1 > t^*$. We therefore write the probability density of ancestor dates under the bounded coalescent model $f(\tau_{1:L-1}|D_{1:K}, \tau_1 > t^*)$ as opposed to $f(\tau_{1:L-1}|D_{1:K})$ for the unbounded coalescent model. The probability density of the ancestor dates under the bounded coalescent model can be rewritten using Bayes rule:

$$f(\tau_{1:L-1}|D_{1:K}, \tau_1 > t^*) = \frac{f(\tau_{1:L-1}|D_{1:K})}{p(\tau_1 > t^*|D_{1:K})} \mathbb{I}(\tau_1 > t^*). \tag{4}$$

Note that the numerator is given by Eq. (3). In other words the bounded coalescent model adds a normalising constant (denominator in Eq. (4)) equal to the probability of all lineages coalescing, $p(\tau_1 > t^*|D_{1:K})$, which we call the "bound probability" and which can be computed as shown in the next section. For two trees satisfying the bound condition, their probability density ratio under the bounded coalescent model is the same as their probability density ratio under the standard (unbounded) coalescent model:

$$\frac{f(\tau\prime_{1:L-1}|D_{1:K}, \tau\prime_1 > t^*)}{f(\tau_{1:L-1}|D_{1:K}, \tau_1 > t^*)} = \frac{f(\tau\prime_{1:L-1}|D_{1:K})}{f(\tau_{1:L-1}|D_{1:K})}. \tag{5}$$

This Eq. (5) clearly follows from Eq. (4) since the denominators are the same and cancel out. Finally, we note that when $t^* \to -\infty$ the bound condition is always satisfied and the bound probability becomes one. In this case the bounded coalescent model reduces to the unbounded coalescent model, which in other words means that the bounded coalescent model is an extension of the unbounded coalescent model.

3. Bound probability

In an isochronous setting the bound probability is equal to the probability that L lineages coalesce into one within the time interval between sampling and t^*. This probability is directly derived from the density function of the time to the MRCA in the standard coalescent model, which can be computed using matrix operations on the Kingman's Markov chain (Tavaré, 1984), using a Laplace transform (Takahata and Nei, 1985) or using a convolution of exponential distributions (Wakeley, 2009) from the coalescent waiting times in Eq. (1). Slightly more generally, the probability that i lineages coalesce down to $j \leqslant i$ lineages in time Δt has also been com-

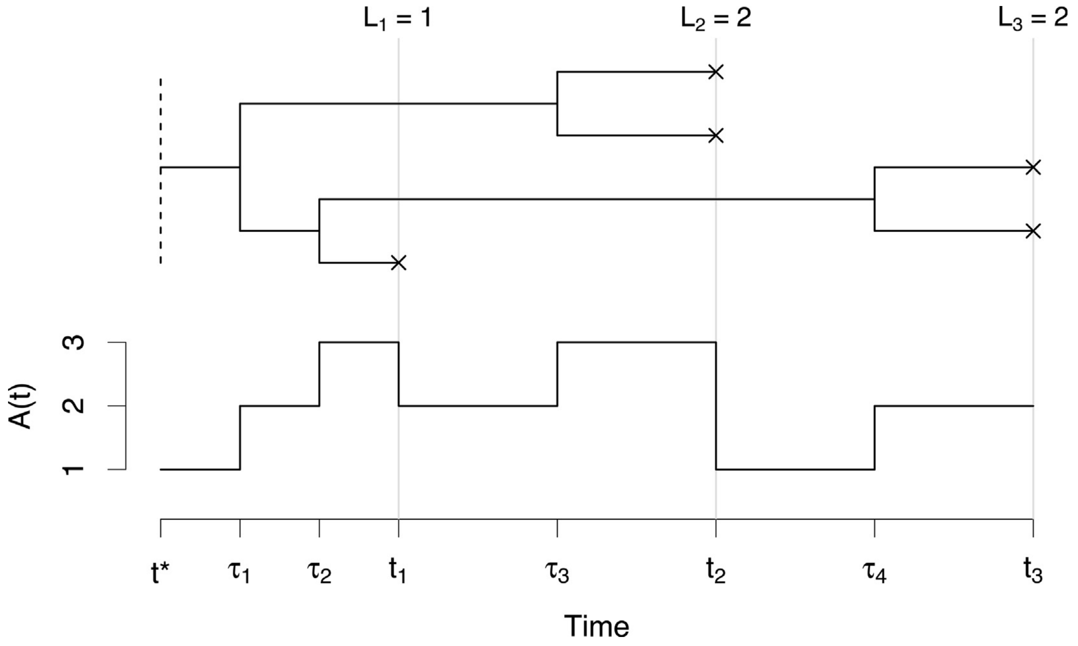


Fig. 1. An example phylogeny in the heterochronous setting. Leaves are taken at times t_1, t_2, and t_3, numbering $L_1 = 1, L_2 = 2$, and $L_3 = 2$ leaves respectively (indicated by ×). Lineages coalesce at times τ_1, τ_2, τ_3, and τ_4. $A(t)$ is the number of lineages that may coalesce at time t. In the bounded coalescent model all lineages must coalesce by time t^* (dashed line), such that $A(t^*) = 1$.

puted before (Tavaré, 1984; Nordborg, 1998; Wakeley, 2009) and can be expressed as

$$g_{i,j}(\Delta t) = \begin{cases} 1 - \left(\sum_{k=2}^{i} e^{-k(k-1)\Delta t/2N_e} \prod_{l=2, l \neq k}^{i} \frac{l(l-1)}{l(l-1)-k(k-1)} \right) & j = 1 \\ \frac{2}{j(j-1)} \sum_{k=j}^{i} \frac{k(k-1)}{2} e^{-k(k-1)\Delta t/2N_e} \prod_{l=j, l \neq k}^{i} \frac{l(l-1)}{l(l-1)-k(k-1)} & j > 1 \end{cases} \tag{6}$$

Hence the bound probability for the isochronous case is simply $p(\tau_1 > t^* | t_1, L) = g_{L,1}(t_1 - t^*)$.

The heterochronous setting is more complex as lineages do not monotonically decrease with every coalescence. To simplify notation, let $A_k = A(t_k)$ denote the number of lineages at time t_k. Note the following Markov property for the unbounded coalescent process

$$p(A_k | A_{k+1:K}) = p(A_k | A_{k+1}). \tag{7}$$

Consequently, we can express the number of lineages at a discrete set of time points as a hidden Markov model (HMM). We can then determine the bound probability using the forward algorithm (Rabiner, 1989; Zucchini and MacDonald, 2009).

The forward algorithm provides the set of probabilities $p(A_k = a_k)$, which are the probabilities of having a_k lineages at time t_k under the standard coalescent model. The algorithm is initialised at time t_K with $p(A_K = L_K) = 1$, and iterates through $k = K-1, K-2, \ldots, 1$ in order to evaluate

$$p(A_k = j + L_k) = \sum_{i=L_{k+1}}^{M_{k+1}} p(A_k = j + L_k | A_{k+1} = i) p(A_{k+1} = i), \tag{8}$$

for $j = 1, \ldots, M_{k+1}$, where $M_k = \sum_{l=k}^{K} L_l$ is the maximum number of lineages that can exist at time t_k. The transition probabilities $p(A_k = j + L_k | A_{k+1} = i)$ correspond to the probability in the standard coalescent model that i lineages coalesce down to j lineages in a time interval $t = t_{k+1} - t_k$ as given in Eq. (6), that is $p(A_k = j + L_k | A_{k+1} = i) = g_{i,j}(t_{k+1} - t_k)$. Since L_k additional leaves are sampled at time t_k, this results in $j + L_k$ lineages.

The forward algorithm is terminated at the bound time t^*. Defining A^* as the number of lineages at the bound time, calculation of the forward probabilities $p(A^*)$ follow Eq. (8) but without new leaves being added:

$$p(A^* = j) = \sum_{i=1}^{L} p(A^* = j | A_1 = i) p(A_1 = i), \tag{9}$$

for $j = 1, \ldots, L$, where $p(A^* = j | A_1 = i) = g_{i,j}(t_1 - t^*)$. The bound probability is then given by $p(A^* = 1)$.

This concludes the calculation of the bound probability in the heterochronous case. This quantity is of interest by itself, but also and perhaps more importantly it allows the calculation of the probability of sampling a tree under the bounded coalescent model by applying Eq. (4). This calculation allows inference to be performed under the bounded coalescent model, either using maximum-likelihood or in a Bayesian framework.

4. Direct sampling

A straightforward approach to simulate realisations of the bounded coalescent model is to use rejection sampling (Didelot et al., 2014; Mallo et al., 2016). This involves simulating from the standard coalescent model and keeping only those simulations in which $\tau_1 > t^*$. This rejection sampling approach can also be used to estimate the bound probability, since acceptance happens with probability equal to the bound probability. However, rejection sampling will be inefficient especially if lineages are sampled close to the bound relative to the effective population size, i.e. if $N_e^{-1}(t_1 - t^*) \gtrsim 0$. In this case the bound probability is small and therefore so is the acceptance probability of the rejection sampler. Here we introduce a direct sampler for the bounded coalescent model that does not suffer this limitation.

The direct sampling approach proceeds through the following five steps:

1. Use the forward filtering backward sampling (FFBS) algorithm to sample the number of lineages $A_{1:K}$ at times $t_{1:K}$ conditioned on the bound.
2. From the sampled $A_{1:K}$ derive the number of coalescence events in the time intervals (t^*, t_1) and (t_k, t_{k+1}) for $k = 1, \dots, K-1$.
3. For intervals containing multiple coalescence events, subdivide the time interval and sample the number of coalescence events for each subinterval. Repeat this step until all $L-1$ coalescence events are constrained by unique, non-overlapping time intervals.
4. Use inverse transform sampling to sample the (constrained) coalescence times.
5. Working backwards in time, sample a pair of lineages for each coalescence event.

Further details of each step are discussed below. We also describe how to calculate the probability density $f(\tau_{1:L-1}|D_{1:K}, \tau_1 > t^*)$ concurrently with the sampling approach, allowing the sampler to be efficiently utilised within an inferential framework. For example it allows the use of the sampler as a proposal distribution in a Markov Chain Monte-Carlo or Importance Sampling algorithm, since in both cases the probability density of the proposed tree would be needed.

4.1. Step 1

In Section 3 we describe how the number of lineages $A^*, A_{1:K}$ can be expressed as a HMM. By treating $A^* = 1$ as an 'observation', we can simulate values of $A_{1:K}$ conditioned on $A^* = 1$ using the FFBS algorithm. FFBS uses a two-step recursion: a forward recursion to calculate the forward probabilities, and a backward recursion to generate samples. The forward recursion is the forward algorithm described in Section 3, so here we focus on the backward recursion for sampling.

The backward recursion is initiated by calculating the probabilities

$$p(A_1 = i|A^* = 1) = \frac{p(A^* = 1|A_1 = i)p(A_1 = i)}{p(A^* = 1)}, \tag{10}$$

for $i = 1, \dots, L$. The terms $p(A_1 = i)$ and $p(A^* = 1)$ are probabilities calculated in the forward algorithm, and the transition probability is given by Eq. (6), i.e. $p(A^* = 1|A_1 = i) = g_{i,1}(t_1 - t^*)$. Once these probabilities have been calculated, the number of lineages a_1 is sampled according to the probabilities $p(A_1 = a_1|A^* = 1)$. The backward recursion then iterates through $k = 2, .., K$, calculating the probabilities

$$p(A_k = i|A_{k-1} = a_{k-1}) = \frac{p(A_{k-1} = a_{k-1}|A_k = i)p(A_k = i)}{p(A_{k-1} = a_{k-1})}, \tag{11}$$

and sampling a value a_k accordingly. Note again that the transition probabilities must account for the addition of leaves, i.e. $p(A_{k-1} = j + L_{k-1}|A_k = i) = g_{i,j}(t_k - t_{k-1})$.

The calculation of $f(\tau_{1:L-1}|D_{1:K}, \tau_1 > t^*)$ is initialised by setting

$$\begin{aligned} \mathscr{F} &:= p(A_{1:K} = a_{1:K}|A^* = 1) \\ &:= p(A_1 = a_1|A^* = 1)\prod_{k=2}^{K} p(A_k = a_k|A_{k-1} = a_{k-1}). \end{aligned} \tag{12}$$

We use $\mathscr{F}$ to emphasise that this is a partial calculation, and will be updated in the following steps.

4.2. Step 2

Having sampled the number of lineages $a_{1:K}$ we derive the number of coalescence events between successive time points. Define $c_{*,1}$ as the number of coalescence events in the interval (t^*, t_1), and $c_{k-1,k}$ as the number of coalescence events in the interval (t_{k-1}, t_k). Then

$$\begin{aligned} c_{*,1} &= a_1 - 1 \\ c_{k-1,k} &= L_{k-1} + a_k - a_{k-1}; \quad k = 2, \dots, K. \end{aligned} \tag{13}$$

4.3. Step 3

In order to sample coalescence times in Step 4, we require that each coalescence event is constrained within a unique, non-overlapping time interval. Hence, for any $c_{k-1,k} \geqslant 2$, we need to partition the corresponding time interval until the coalescence events are separated. Here, we bisect the interval (t_{k-1}, t_k) and sample the number of coalescent events in the subintervals $(t_{k-1}, t_{k-0.5})$ and $(t_{k-0.5}, t_k)$, where $t_{k-0.5} = 0.5(t_{k-1} + t_k)$. This is achieved by sampling the number of lineages $a_{k-0.5}$ at the newly added time point $t_{k-0.5}$ according to

$$\begin{aligned} &p(A_{k-0.5} = a_{k-0.5}|A_{k-1} = a_{k-1}, A_k = a_k) = \\ &\tfrac{p(A_{k-1}=a_{k-1}|A_{k-0.5}=a_{k-0.5})p(A_{k-0.5}=a_{k-0.5}|A_k=a_k)}{p(A_{k-1}=a_{k-1}|A_k=a_k)}, \end{aligned} \tag{14}$$

and then deriving the number of coalescence events $c_{k-1,k-0.5}$ and $c_{k-0.5,k}$ as in Step 2. The probability of sampling $a_{k-0.5}$ is incorporated into the calculation of $f(\tau_{1:L-1}|D_{1:K}, \tau_1 > t^*)$:

$$\mathscr{F} := \mathscr{F} \times p(A_{k-0.5} = a_{k-0.5}|A_{k-1} = a_{k-1}, A_k = a_k). \tag{15}$$

This process is also applied to the interval (t^*, t_1) if $c_{*,1} \geqslant 2$, and any newly generated subintervals until we have at most one coalescence event in any defined time interval.

4.4. Step 4

Each coalescence event is constrained by a unique non-overlapping time interval, but is not uniformly distributed within said interval. From the previous steps, assume that a coalescence event occurs in the interval (t_{k-1}, t_k), with a_k lineages at time t_k and $a_{k-1} = a_k - 1$ lineages at time t_{k-1}. The probability density of the coalescence time τ is proportional to the density of the coalescence time under the standard coalescent model multiplied by the probability that no further coalescence events occur in the interval (t_{k-1}, τ), giving

$$\begin{aligned} f(\tau|t_{k-1}, t_k, a_k) \propto &\frac{a_k(a_k-1)}{2N_e} \exp\left(-\frac{a_k(a_k-1)}{2N_e}(t_k - \tau)\right) \\ &\times \exp\left(-\frac{(a_k-1)(a_k-2)}{2N_e}(\tau - t_{k-1})\right), \end{aligned} \tag{16}$$

for $t_{k-1} < \tau < t_k$. This enables us to sample τ using inverse transform sampling. Collecting the τ terms of Eq. (16) gives

$$f(\tau|t_{k-1}, t_k, a_k) = \frac{1}{Z} \exp\left(\frac{a_k - 1}{N_e}\tau\right), \quad t_{k-1} < \tau < t_k, \tag{17}$$

where Z is the normalising constant, and is equal to

$$Z = \frac{N_e}{a_k - 1}\left(\exp\left(\frac{a_k - 1}{N_e}t_k\right) - \exp\left(\frac{a_k - 1}{N_e}t_{k-1}\right)\right). \tag{18}$$

If u is a draw from a standard uniform distribution, we can compute τ using inverse transform sampling by solving

$$u = \frac{N_e}{Z(a_k - 1)}\left(\exp\left(\frac{a_k - 1}{N_e}t_k\right) - \exp\left(\frac{a_k - 1}{N_e}\tau\right)\right), \tag{19}$$

where the right side is the cumulative density function of τ. This gives

$$\tau = \frac{N_e}{a_k - 1}\log\left(\exp\left(\frac{a_k - 1}{N_e}t_k\right) - \frac{a_k - 1}{N_e}Zu\right). \tag{20}$$

Each time a coalescence time is sampled, the calculation of $f(\tau_{1:L-1}|D_{1:K},\tau_1 > t^*)$ is updated using

$$\mathscr{F} := \mathscr{F} \times f(\tau|t_{k-1},t_k,a_k). \tag{21}$$

By using inverse transform sampling for each coalescence event, we obtain the full set of coalescence times $\tau_{1:L-1}$. The probability density $\mathscr{F}$ at the end of this step is equal to that given by Eq. (4), i.e. $\mathscr{F} = f(\tau_{1:L-1}|D_{1:K},\tau_1 > t^*)$.

4.5. Step 5

The final step is to sample the topology of the tree. Here we simply iterate backwards in time through the coalescence events and sample two of the extant lineages, noting the ancestors for each coalescent node. The probability of a topology G conditional on the sampled $\tau_{1:L-1}$ is given by

$$p(G|\tau_{1:L-1},D_{1:K}) = \prod_{i=1}^{L-1}\frac{2}{A(\tau_i)(A(\tau_i)-1)}. \tag{22}$$

The density of the sampled tree under the bounded coalescent model is then given by

$$f(G,\tau_{1:L-1}|D_{1:K},\tau_1 > t^*) = p(G|\tau_{1:L-1},D_{1:K}) \times f(\tau_{1:L-1}|D_{1:K},\tau_1 > t^*). \tag{23}$$

4.6. Validation and comparison with rejection sampling

4.6.1. Bound probability

Here we consider an example, with $L = 5$ leaves at different times $t_1 = 0.0, t_2 = 0.5, t_3 = 1.0, t_4 = 1.5, t_5 = 2.0$ and a bound time $t^* = -0.5$. We use an effective population size of $N_e = 1$ so that pairs of lineages coalesce at rate $1/N_e = 1$. We first used a rejection approach to simulate under these conditions. The rejection sampler required 427 371 simulations in order to obtain 10^5 acceptances, which means that the bound probability is estimated to be 0.234.

Next we run the forward filter described in Section 3, giving the probabilities shown in Table 1. In particular we note that the bound probability of having a single lineage at the bound time t^* is 0.233, which is consistent with the bound probability estimated by the rejection sampler.

4.6.2. A single run of the direct sampler

Step 1 of the direct sampler consists in running the forward algorithm to obtain the probabilities shown in Table 1, followed by backwards sampling conditioned on the bound, that is $A^* = 1$. We update the filtered probabilities for A_1 using Eq. (11), giving the probability vector $(0.000, 0.411, 0.494, 0.093, 0.002)^\intercal$, from which we sample. Say we sample $A_1 = 4$, we iterate to time t_2 and obtain the probabilities $p(A_2 = 3|A_1 = 4) = 0.736$ and $p(A_2 = 4|A_1 = 4) = 0.264$. Note that $p(A_2 = 2|A_1 = 4) = 0$ since 2 leaves at time t_2 can not result in 4 leaves at time t_1. This sampling procedure continues until we have samples $a_1,\ldots,a_5$.

Table 1
Forward filter probabilities in the numerical example (probabilities may not add to one due to rounding).

Lineages	$t^* = -0.5$	$t_1 = 0.0$	$t_2 = 0.5$	$t_3 = 1.0$	$t_4 = 1.5$	$t_5 = 2.0$
1	0.233	0	0	0	0	1
2	0.565	0.244	0.277	0.393	1	0
3	0.192	0.571	0.587	0.607	0	0
4	0.010	0.178	0.135	0	0	0
5	0.000	0.007	0	0	0	0

In Step 2 we determine the number of coalescence events between time points. If in Step 1 we sample $a_1 = 4, a_2 = 3, a_3 = 3, a_4 = 2, a_5 = 1$ we can determine that $c_{*,1} = 3$ coalescence events occur in the interval (t^*,t_1), and $c_{2,3} = 1$ in (t_2,t_3).

In Step 3 we further partition the time axis in order to separate each coalescence event. Here we would add a new time point at $t_{0.5} = -0.25$, and sample the extant number of lineages conditional on having a single lineage at time t^* and four lineages at time t_1. Hence we may sample 0, 1, 2, or 3 coalescence events in the interval $(t^*,t_{0.5})$, with the remainder occurring in the interval $(t_{0.5},t_1)$. This partitioning continues until each coalescence events is constrained by a unique non-overlapping time interval, i.e. $(-0.5,-0.25),(-0.25,-0.125),(-0.125,0.0),(0.5,1.0)$.

Step 4 is simply a matter of sampling the coalescence times, conditioned on the corresponding intervals.

Finally, in Step 5 we sample the lineages for each coalescence event. Starting with the latest event, we determine which lineages currently exist. For the interval $(0.5,1.0)$ these would be the lineages corresponding to leaves 3, 4, and 5. In the interval $(-0.125,0.0)$, these would be the lineages corresponding to leaves 1 and 2, as well as the resulting lineages from the previous coalescence event.

4.6.3. Comparison of simulated trees using direct and rejection sampling

We obtain 10^5 simulations from both the rejection sampling algorithm and the direct sampling algorithm as described above. In Fig. 2 we compare the simulated coalescence times, and observe strong agreement between the two methods. This example also demonstrates complex behaviour arising from the heterochronous setting. In particular, the distributions of the latter coalescence times are multimodal. This results from lineages not coalescing by the time a new leaf is added, leading to an increased coalescent rate.

We also compare the run times of the two algorithms as we vary the number of leaves. Keeping $t^* = -0.5$ and $N_e = 1$ fixed, we sample L leaves uniformly over the interval $(0,2)$ and obtain 10^5 simulations from each algorithm. The run times for $L = 2,\ldots,50$ are shown in Fig. 3. For a small number of leaves, both algorithms exhibit similar performance. However, as the number of leaves increases, the computational cost of rejection sampling increases much more rapidly than the direct sampling approach. This results from the smaller bound probabilities lowering the acceptance rate of the rejection sampler.

5. Properties

In this section we investigate the effect of the bound on various properties of the tree, which can differ markedly from the standard coalescent model and alter the outcome of many standard analyses.

5.1. Dependencies between coalescent events

The bound induces dependencies between the waiting times of the coalescence events. As a demonstration, we consider the isochronous setting with $L = 3, t_1 = 0$, and estimate the correlation between the two waiting times for bound times $-5 \leqslant t^* \leqslant -0.01$. This estimation is performed numerically by simulating 10^5 trees for each bound time. The results are shown in Fig. 4. As the bound time increases, the correlation between waiting times becomes stronger. Since all lineages must coalesce by the bound time, if one waiting time is large then the other must be small. As the bound time decreases and the bounded coalescent model more

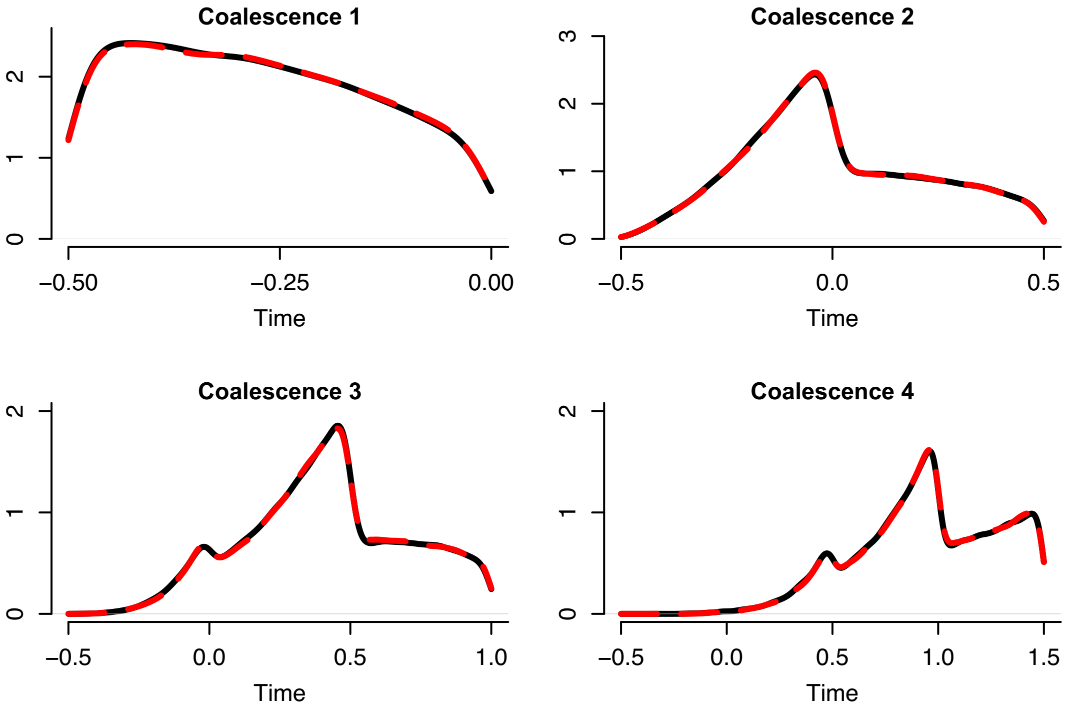


Fig. 2. Kernel density estimates of the four coalescence times in the heterochronous setting with $L = 5$ leaves at times $t_1 = 0.0, t_2 = 0.5, t_3 = 1.0, t_4 = 1.5, t_5 = 2.0$, effective population size $N_e = 1$, and bound time $t^* = -0.5$. The solid black line is the density obtained using direct sampling, and the dashed red line is the density obtained using rejection sampling.

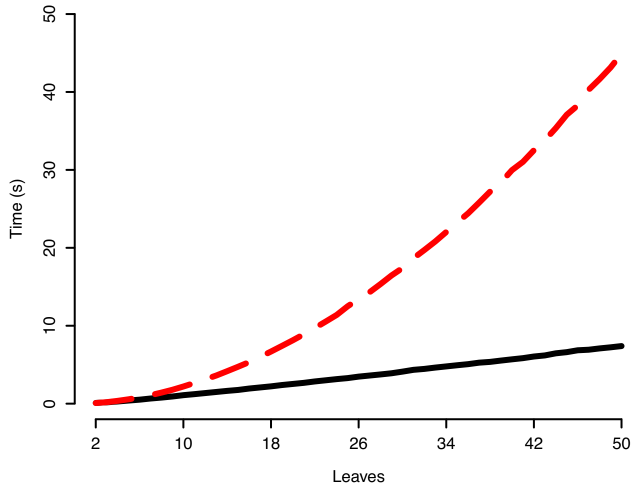


Fig. 3. Comparison of run times for direct sampling and rejection sampling with varying number of leaves sampled over the interval $(0, 2)$ with bound time $t^* = -0.5$. The solid black line is the run time for direct sampling, and the dashed red line is the run time for rejection sampling. 10^5 simulations are obtained in each instance.

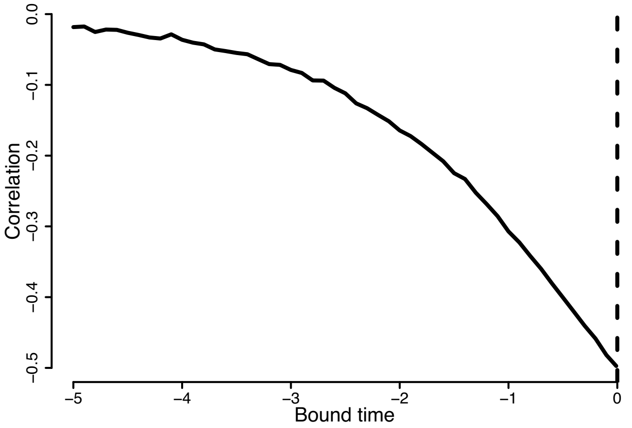


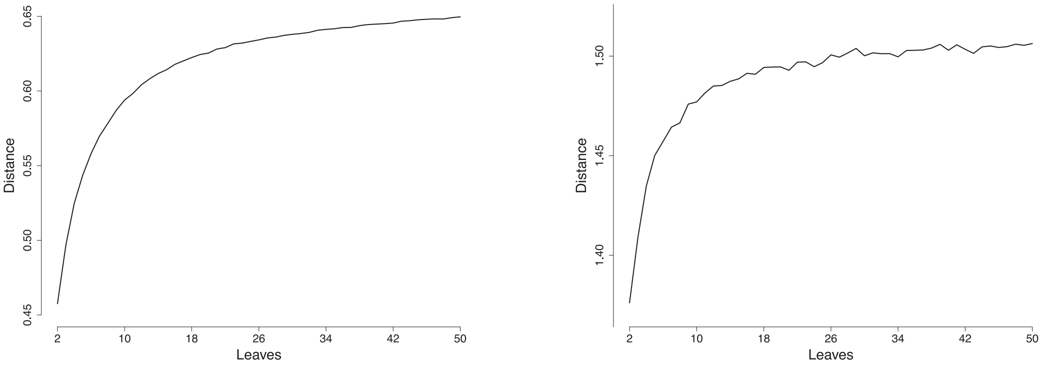
Fig. 4. Correlation between the two waiting times when $L = 3, t_1 = 0$ (dashed line) and $N_e = 1$.

strongly approximates the standard coalescent model, the correlation tends towards zero.

Due to these dependencies, it is necessary to consider all leaves and lineages jointly when performing simulations or deriving probabilities under the bounded coalescent model. For instance, simulation under the standard coalescent model can be undertaken by adding leaves one by one to the growing tree, but this is not the case under the bounded coalescent model. To illustrate, we again consider the isochronous setting and estimate by simulation the average pairwise distance between leaves for $L = 2, \ldots, 50$, while keeping t^* fixed at -0.5 (Fig. 5)a and and -2 (Fig. 5b). In both cases, increasing the total number of leaves increases their average pairwise distance. Consequently, simulating a tree under the bounded coalescent model with $L = 3$ can not be achieved by first simulating a tree under the bounded coalescent model with $L = 2$ and then adding an extra leaf. Rather, the initial two leaves would need to be simulated conditional on all three leaves coalescing by the bound time.

5.2. Bound probability

For further exploration we consider the heterochronous setting where L leaves are sampled at evenly spaced times over the sampling interval $(0, t_L)$. The isochronous setting can be recovered by setting $t_L = 0$. We investigate how properties of the simulated trees change as L, t_L, and t^* are varied.



(a) $t^* = -0.5$

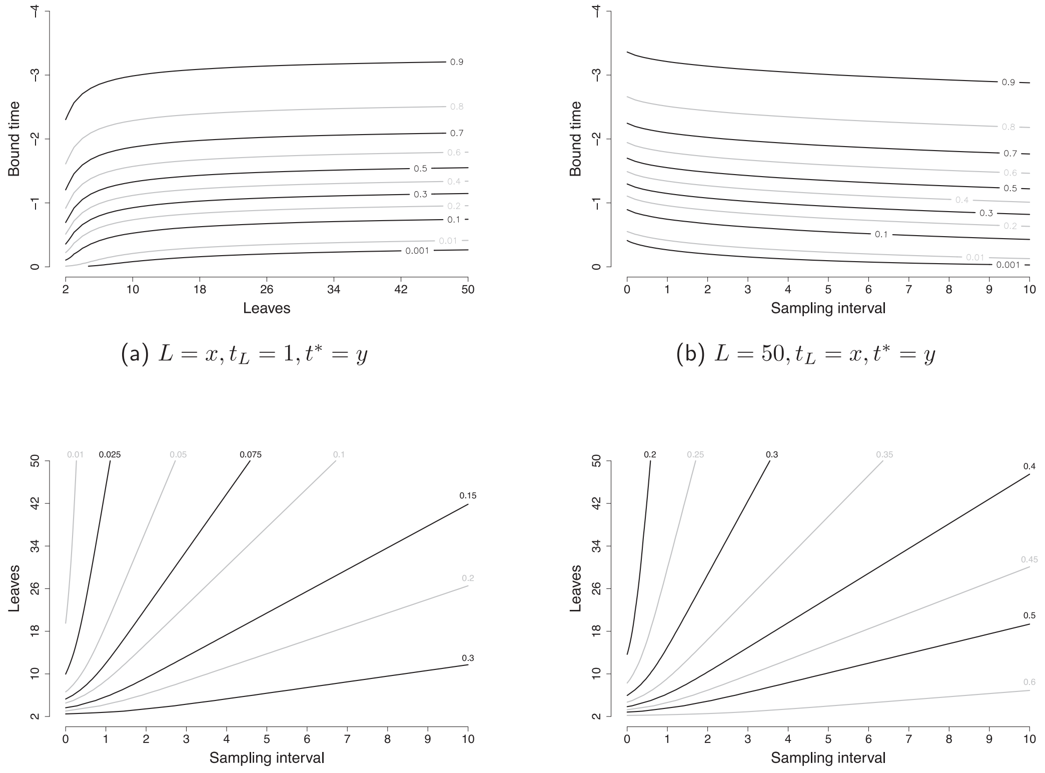
(b) $t^* = -2$

Fig. 5. Average distance between two leaves for different values of L keeping $t_1 = 0$ and $N_e = 1$ fixed.

Estimates of the bound probability for varying L, t_L, and t^* are shown in Fig. 6. The value of the bound time t^* has the largest impact. As the bound time moves further into the past the bound probability tends towards one, even for large numbers of leaves and/or short sampling intervals. Sampled trees will strongly resemble those obtained under the standard coalescent model. On the other hand, as the bound time becomes more recent, the bound probability tends towards zero. In this region sampled trees will have very different properties than the standard coalescent model. For modest bound times, the choice of both L and t_L significantly impact the bound probability, which decreases for increasing L and decreasing t_L. The largest changes are observed for small L and t_L, and asymptotic behaviour is observed for large L and t_L.

5.3. Tree summary statistics

Fig. 7 shows how several summary statistics of bounded coalescent trees change as the bound time is altered. In particular we consider the distributions of the time of the MRCA (TMRCA), the total branch length, the average pairwise distance between the leaves, and the ratio between the average terminal branch length to the average internal branch length (starlikeness). The total number of leaves is fixed at $L = 50$ in Fig. 7, and we consider three sampling intervals, $t_L = 0$ (isochronous), $t_L = 1$, and $t_L = 10$. Within these three configurations, we consider the four bound times $t^* = -\infty, t^* = -2, t^* = -1$, and $t^* = -0.5$, and sample 1000 trees in each of the twelve cases to approximate the resulting distributions.



(a) $L = x, t_L = 1, t^* = y$

(b) $L = 50, t_L = x, t^* = y$

(c) $L = y, t_L = x, t^* = -0.5$

(d) $L = y, t_L = x, t^* = -1$

Fig. 6. Bound probabilities as the bound time and number of leaves vary in the heterochronous example. (a) uses a sampling interval of $t_L = 1$, (b) uses $L = 50$ leaves, (c) uses a bound time of $t^* = -0.5$, and (d) uses a bound time of $t^* = -1$. The remaining parameters in each simulation are given by the x-axis (x) and y-axis (y).

Since all lineages are required to coalesce before the bound time, providing a lower bound for TMRCA, the bounded coalescent model leads to more recent values of TMRCA. As the bound time increases, values of TMRCA become increasingly concentrated near the bound time. Increasing the bound time tends to reduce the total branch length. This results from requiring all lineages to coalesce by the bound time, which gives an upper limit of $((t_L/2) - t^*)L$.

As with the total branch length, the bounded coalescent model places an upper bound on the average pairwise distance, which is $(L+1)/(3(L-1))$. This causes the average pairwise distance to decrease as the bound time increases. Finally, the starlikeness increases as the bound time increases.

5.4. Phylodynamics

The starlikeness of a tree is often an indication of past population size growth (Slatkin and Hudson, 1991; den Bakker et al., 2008; Volz et al., 2009). Therefore, the fact that this tree summary statistic depends on the bound time (Fig. 7) suggests that the pres-

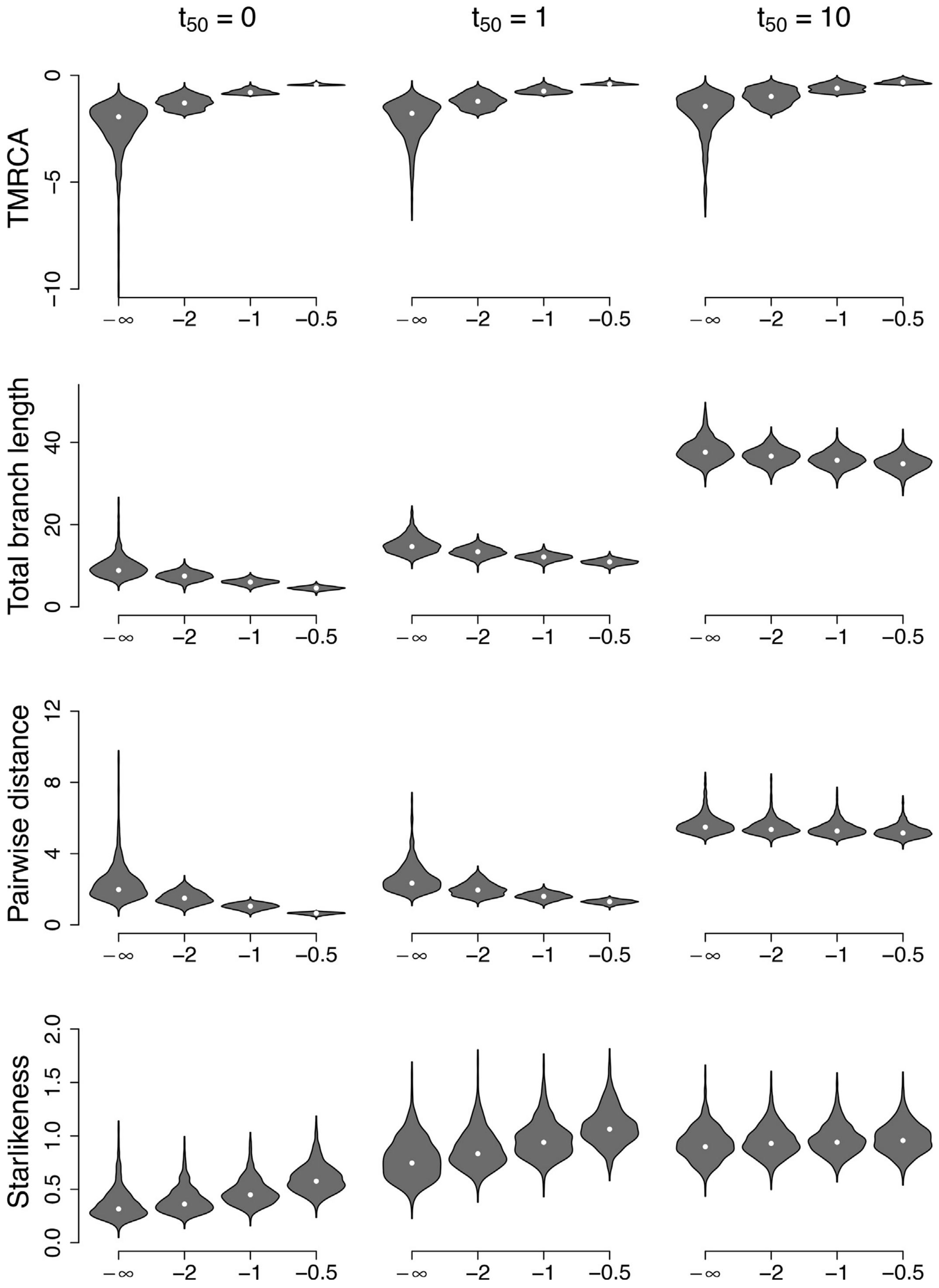


Fig. 7. Properties of the bounded coalescent model with $L = 50$ leaves, sampling intervals of length $t_{50} = 0$ (isochronous), $t_{50} = 1$, and $t_{50} = 10$, and bound time t^* of $-\infty$ (corresponding to the standard unbounded coalescent), $-2, -1$ and -0.5. Mean values are shown as white circles.

ence of a bound could confound phylodynamic inference studies, which are aimed at reconstruct past population size dynamics given genetic or phylogenetic data (Nee et al., 1995; Pybus et al., 2000; Ho and Shapiro, 2011).

To illustrate this, Fig. 8 shows two examples of skyline plots in the isochronous setting and two examples in the heterochronous setting. All skyline plots were computed using Bayesian nonparametric phylodynamic reconstruction (Palacios and Minin, 2012) as implemented in the R package phylodyn (Karcher et al., 2017). In all four examples the population size was incorrectly inferred to have grown significantly. In examples (a) and (c) this was caused by a high effective population size relative to the bound time, so that the bound conditions forces coalescence to happen before it would normally do, whereas in examples (b) and (d) the same occurred due to relatively recent bound times. Similarly, a previous study showed that having a bound mimics the effect of population growth on the site frequency spectrum (Lapierre et al., 2017), even though the two processes are not equivalent. These results do not invalidate the principles of phylodynamic inference, which assume that there is no bound on the root date, but warn against its application in situations where a bound is present.

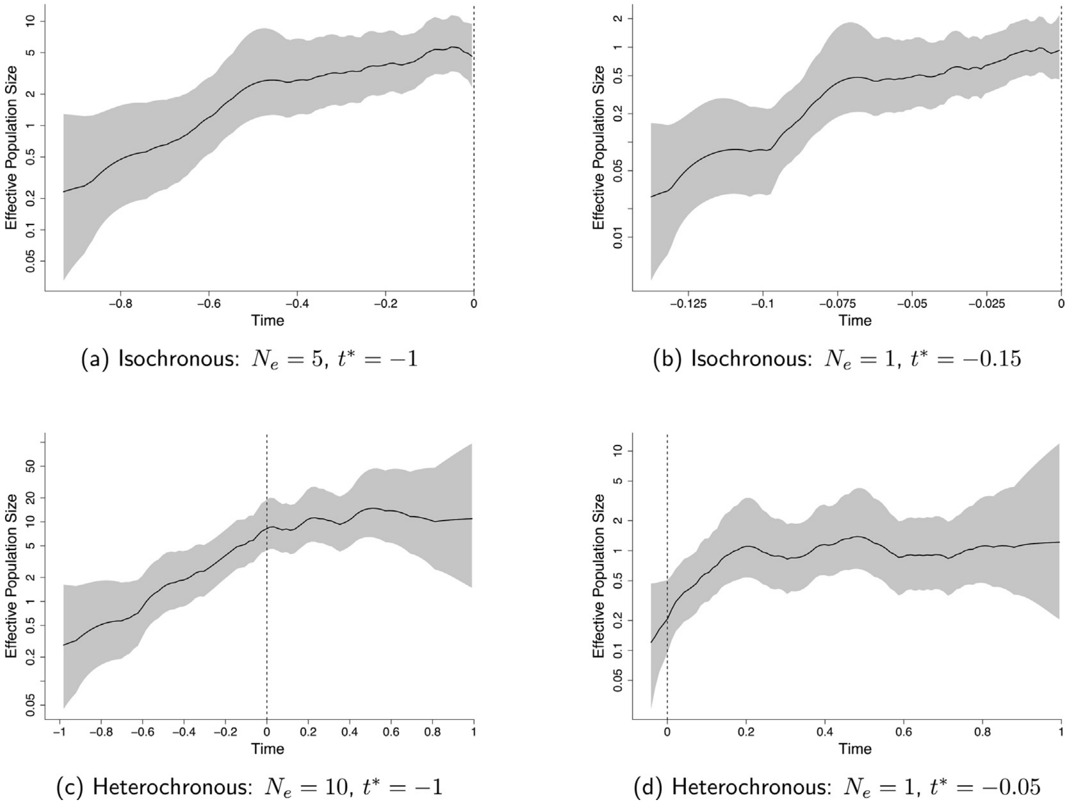
6. Implementation

We implemented the algorithms and methods described in this paper into a new R package called *BoundedCoalescent* which is available at https://github.com/DrJCarson/BoundedCoalescent. This package includes functions to calculate the bound probability, to calculate the probability density of a tree under the bounded coalescent model, and to simulate trees under the bounded coalescent model. Most of the code was written in C++ and integrated into the R package using Rcpp (Eddelbuettel and François, 2011; Eddelbuettel, 2013). The R package ape was used to store, manipulate and visualise phylogenetic trees (Paradis and Schliep, 2019).

7. Discussion

In this paper we have presented a formal description of the bounded coalescent model (Rasmussen and Kellis, 2012), an extension of the standard coalescent model in which all lineages are constrained to find a common ancestor by a predefined date. We have shown how to calculate the probability of the bound constraint happening by chance, which is useful to calculate the probability density of a given phylogeny under the bounded coalescent model with a given bound time. We have also described a method to directly sample phylogenies under the bounded coalescent model, and used this to explore the properties of the model and the effect of the conditioning. Although we focused on the case of a lower bound on the root date, we note that our results can also be used in situations where the root date has an upper bound, for example due to fossil records (Ho et al., 2014). The probability density of a tree in this case can be obtained by dividing the unbounded probability density by one minus the bound probability, similar to Eq. (4). Sampling can be achieved by first selecting a number of existing lineages of at least two at the bound time, and simulating the constrained coalescent process after the bound time using the same constrained procedure as described in Section 4. The only additional computation is to also simulate the coalescent process before the bound time, which simply follows the unconstrained coalescent process (Eq. (1)).

The standard coalescent framework can be extended in many ways (Donnelly and Tavare, 1995; Fu and Li, 1999; Rosenberg



(a) Isochronous: $N_e = 5,\ t^* = -1$

(b) Isochronous: $N_e = 1,\ t^* = -0.15$

(c) Heterochronous: $N_e = 10,\ t^* = -1$

(d) Heterochronous: $N_e = 1,\ t^* = -0.05$

Fig. 8. Skyline plots with $L = 50$. The solid lines show the median population estimates, and the grey regions show 95% credible intervals. The plotted range indicates the bound time and the latest leaf. The dashed line shows the earliest sampled leaf.

and Nordborg, 2002), for example to allow for variations in the population size (Griffiths and Tavare, 1994), geographical structure (Notohara, 1990), recombination (Hudson, 1990) and selection (Krone and Neuhauser, 1997). All of these extensions are in principle compatible with the conditioning imposed by the bounded coalescent model. For example, an ancestral recombination graph could be simulated, for which efficient methods have been developed (McVean and Cardin, 2005), and rejection sampling could be applied to ensure that the bound condition is met. However, it is unclear under which of these coalescent framework extensions a direct sampling method can be devised or an efficient method to calculate the probability density of realisations. As for other extensions of the coalescent framework, there are data analysis situations where it is unclear whether a bounded model should be used or not. Having a well-defined bounded coalescent model including an algorithm for computing tree probability densities allows for model selection techniques to be used in such situations (Xie et al., 2011).

The bounded coalescent model is of special interest for applying coalescent theory in infectious disease epidemiology. This includes the need to have full coalescence of lineages within a host before that host become infected, if we assume a complete transmission bottleneck (Didelot et al., 2014; Didelot et al., 2017). The effect of a complete transmission bottleneck becomes more important as data on within–host diversity are increasingly being used to infer who infected whom (De Maio et al., 2018; Wymant et al., 2018). An alternative approach to the bounded coalescent model in order to enforce a complete transmission bottleneck is to start the within–host population with an effective population size of zero, with subsequent growth, so that the coalescence rate is close to infinity just after infection. For example, a within–host linear growth model starting at zero was used as part of a method for simultaneous inference of phylogenetic and transmission trees (Klinkenberg et al., 2017). Similarly, a recently proposed model on clonal expansion has each expansion starting with a size of zero to ensure they initially correspond to a single lineage (Helekal et al., 2021). Finally, the coalescent process is sometimes equated with the transmission process by assuming that the within–host population size is negligible, so that coalescent times correspond to transmission events (Volz et al., 2009; Frost and Volz, 2010). In this framework, considering that incidence is proportional to prevalence, as is the case in many infectious disease epidemiology models, leads to an effective population size proportional to the number of infected individuals minus one (Volz, 2012; Volz and Didelot, 2018). Consequently, the bound condition is enforced by having an effective population size of zero at the time when an outbreak is seeded with a single index case.

CRediT authorship contribution statement

Jake Carson: Conceptualization, Methodology, Software, Investigation, Writing - original draft, Writing - review & editing. **Alice Ledda:** Conceptualization, Investigation, Writing - review & editing. **Luca Ferretti:** Conceptualization, Investigation, Writing - review & editing. **Matt Keeling:** Conceptualization, Investigation, Writing - review & editing, Supervision, Funding acquisition. **Xavier Didelot:** Conceptualization, Methodology, Investigation, Investigation, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge funding from the National Institute for Health Research (NIHR) Health Protection Research Unit in Genomics and Enabling Data (Grant No. NIHR200892).

References

den Bakker, H.C., Didelot, X., Fortes, E.D., Nightingale, K.K., Wiedmann, M., 2008. Lineage specific recombination rates and microevolution in Listeria monocytogenes. BMC Evol. Biol. 8, 277. https://doi.org/10.1186/1471-2148-8-277.

Cannings, C., 1974. The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. Adv. Appl. Probab. 6, 260–290. https://doi.org/10.2307/1426293.

De Maio, N., Worby, C.J., Wilson, D.J., Stoesser, N., 2018. Bayesian reconstruction of transmission within outbreaks using genomic variants. PLOS Comput. Biol. 14,. https://doi.org/10.1371/journal.pcbi.1006117 e1006117.

Didelot, X., Fraser, C., Gardy, J., Colijn, C., 2017. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Mol. Biol. Evol. 34, 997–1007. https://doi.org/10.1093/molbev/msw275.

Didelot, X., Gardy, J., Colijn, C., 2014. Bayesian inference of infectious disease transmission from whole-genome sequence data. Mol. Biol. Evol. 31, 1869–1879. https://doi.org/10.1093/molbev/msu121.

Donnelly, P., Tavare, S., 1995. Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. 29, 401–421. https://doi.org/10.1146/annurev.ge.29.120195.002153.

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G., Solomon, W., 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161, 1307–1320. https://doi.org/10.1093/genetics/161.3.1307.

Drummond, A.J., Pybus, O.G., Rambaut, A., Forsberg, R., Rodrigo, A.G., 2003. Measurably evolving populations. Trends Ecol. Evol. 18, 481–488. https://doi.org/10.1016/S0169-5347(03)00216-7.

Du, P., Ogilvie, H.A., Nakhleh, L., 2019. Unifying gene duplication, loss, and coalescence on phylogenetic networks. In: Lect. Notes Comput. Sci.. Springer International Publishing. volume 11490 LNBI, pp. 40–51. https://doi.org/10.1007/978-3-030-20242-2_4.

Eddelbuettel, D., 2013. Seamless R and C++ integration with Rcpp. Springer, New York. https://doi.org/10.1007/978-1-4614-6868-4.

Eddelbuettel, D., François, R., 2011. Rcpp: seamless R and C++ integration. J. Stat. Softw. 40, 1–18. https://doi.org/10.18637/jss.v040.i08.

Ferretti, L., Disanto, F., Wiehe, T., 2013. The effect of single recombination events on coalescent tree height and shape. PLoS One 8,. https://doi.org/10.1371/journal.pone.0060123 e60123.

Fisher, R.A., 1930. The genetical theory of natural selection. Clarendon Press. https://doi.org/10.5962/bhl.title.27468.

Frost, S.D.W., Volz, E.M., 2010. Viral phylodynamics and the search for an 'effective number of infections'. Philos. Trans. R. Soc. B 365, 1879–1890. https://doi.org/10.1098/rstb.2010.0060.

Fu, Y.X., Li, W.H., 1999. Coalescing into the 21st century: An overview and prospects of coalescent theory. Theor. Popul. Biol. 56, 1–10. https://doi.org/10.1006/tpbi.1999.1421.

Griffiths, R.C., Tavare, S., 1994. Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. B 344, 403–410. https://doi.org/10.1098/rstb.1994.0079.

Helekal, D., Ledda, A., Volz, E., Wyllie, D., Didelot, X., 2021. Bayesian inference of clonal expansions in a dated phylogeny. Syst. Bio. https://doi.org/10.1093/sysbio/syab095. TBD, syab095.

Hill, M., Legried, B., Roch, S., 2020. Species tree estimation under joint modeling of coalescence and duplication: sample complexity of quartet methods. arXiv, 2007.06697..

Ho, S.Y.W., Duchêne, S., 2014. Molecular-clock methods for estimating evolutionary rates and timescales. Mol. Ecol. 23, 5947–5965. http://doi.wiley.com/10.1111/mec.12953, 10.1111/mec.12953..

Ho, S.Y.W., Shapiro, B., 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. Mol. Ecol. Resour. 11, 423–434. https://doi.org/10.1111/j.1755-0998.2011.02988.x.

Hudson, R.R., 1990. Gene genealogies and the coalescent process. Oxford Surv. Evol. Biol. 7, 1–44.

Karcher, M.D., Palacios, J.A., Lan, S., Minin, V.N., 2017. PHYLODYN: an R package for phylodynamic simulation and inference. Mol. Ecol. Resour. 17, 96–100. https://doi.org/10.1111/1755-0998.12630.

Kingman, J.F.C., 1982. On the genealogy of large populations. J. Appl. Probab. 19, 27–43. https://doi.org/10.2307/3213548.

Kingman, J.F.C., 1982. The coalescent. Stoch. Process. their Appl. 13, 235–248. https://doi.org/10.1016/0304-4149(82)90011-4.

Klinkenberg, D., Backer, J.A., Didelot, X., Colijn, C., Wallinga, J., 2017. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLoS Comput. Biol. 13,. https://doi.org/10.1371/journal.pcbi.1005495 e1005495.

Krone, S.M., Neuhauser, C., 1997. Ancestral processes with selection. Theor. Popul. Biol. 51, 210–237. https://doi.org/10.1006/tpbi.1997.1299.

Lapierre, M., Lambert, A., Achaz, G., 2017. Accuracy of demographic inferences from the site frequency spectrum: The case of the yoruba population. Genetics 206, 139–449. https://doi.org/10.1534/genetics.116.192708.

Li, Q., Scornavacca, C., Galtier, N., Chan, Y.B., 2021. The multilocus multispecies coalescent: a flexible new model of gene family evolution. Syst. Biol. 70, 822–837. https://doi.org/10.1093/sysbio/syaa084.

Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523–536. https://doi.org/10.1093/sysbio/46.3.523.

Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21–30. https://doi.org/10.1080/10635150500354928.

Mallo, D., de Oliveira Martins, L., Posada, D., 2016. SimPhy: phylogenomic simulation of gene, locus and species trees. Syst. Biol. 65, 334–344. https://doi.org/10.1093/sysbio/syv082.

McVean, G.A.T., Cardin, N.J., 2005. Approximating the coalescent with recombination. Phil. Trans. R. Soc. B 360, 1387–1393. https://doi.org/10.1098/rstb.2005.1673.

Moran, P.A.P., 1958. Random processes in genetics. Math. Proc. Cambridge Philos. Soc. 54, 60–71. https://doi.org/10.1017/S0305004100033193.

Nee, S., Holmes, E.C., Rambaut, A., Harvey, P.H., 1995. Inferring population history from molecular phylogenies. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 349, 25–31. https://doi.org/10.1098/rstb.1995.0087.

Nordborg, M., 1998. On the probability of Neanderthal ancestry. Am. J. Hum. Genet. 63, 1237–1240. https://doi.org/10.1086/302052.

Notohara, M., 1990. The coalescent and the genealogical process in geographically structured population. J. Math. Biol. 29, 59–75. https://doi.org/10.1007/BF00173909.

Palacios, J.A., Minin, V.N., 2012. Integrated nested Laplace approximation for Bayesian nonparametric phylodynamics, in: Uncertain. Artif. Intell. - Proc. 28th Conf. UAI 2012, pp. 726–735..

Paradis, E., Schliep, K., 2019. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35, 526–528. https://doi.org/10.1093/bioinformatics/bty633.

Pybus, O.G., Rambaut, A., Harvey, P.H., 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 155, 1429–1437. https://doi.org/10.1093/genetics/155.3.1429.

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77, 257–286. https://doi.org/10.1109/5.18626.

Rambaut, A., 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 16, 395–399. https://doi.org/10.1093/bioinformatics/16.4.395.

Rasmussen, M.D., Hubisz, M.J., Gronau, I., Siepel, A., 2014. Genome-wide inference of ancestral recombination graphs. PLoS Genet. 10. https://doi.org/10.1371/journal.pgen.1004342.

Rasmussen, M.D., Kellis, M., 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res. 22, 755–765. https://doi.org/10.1101/gr.123901.111.

Rosenberg, N.A., Nordborg, M., 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. 3, 380–390. https://doi.org/10.1038/nrg795.

Slatkin, M., Hudson, R.R., 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129, 555–562. https://doi.org/10.1093/genetics/129.2.555.

Takahata, N., Nei, M., 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110, 325–344. https://doi.org/10.1093/genetics/110.2.325.

Tavaré, S., 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26, 119–164. https://doi.org/10.1016/0040-5809(84)90027-3.

Volz, E.M., 2012. Complex population dynamics and the coalescent under neutrality. Genetics 190, 187–201. https://doi.org/10.1534/genetics.111.134627.

Volz, E.M., Didelot, X., 2018. Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance. Syst. Biol. 67, 719–728. https://doi.org/10.1093/sysbio/syy007.

Volz, E.M., Kosakovsky Pond, S.L., Ward, M.J., Leigh Brown, A.J., Frost, S.D.W., 2009. Phylodynamics of infectious disease epidemics. Genetics 183, 1421–1430. https://doi.org/10.1534/genetics.109.106021.

Wakeley, J., 2009. Coalescent theory: an introduction. Roberts and Company Publishers.

Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159. https://doi.org/10.1093/genetics/16.2.97.

Wymant, C., Hall, M., Ratmann, O., Bonsall, D., Golubchik, T., de Cesare, M., Gall, A., Cornelissen, M., Fraser, C., 2018. PHYLOSCANNER: Inferring transmission from within- and between-host pathogen genetic diversity. Mol. Biol. Evol. 35, 719–733. https://doi.org/10.1093/molbev/msx304.

Xie, W., Lewis, P.O., Fan, Y., Kuo, L., Chen, M.H., 2011. Improving marginal likelihood estimation for bayesian phylogenetic model selection. Syst. Biol. 60, 150–160. https://doi.org/10.1093/sysbio/syq085.

Zucchini, W., MacDonald, I.L., 2009. Hidden Markov models for time series: an introduction using R. Chapman and Hall/CRC.