Coalescent models derived from birth–death processes

Crespo, Fausto F.; Posada, David; Wiuf, Carsten

Published in:
Theoretical Population Biology

DOI:
10.1016/j.tpb.2021.09.003

Publication date:
2021

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Coalescent models derived from birth–death processes

Fausto F. Crespo a,b, David Posada a,b,c, Carsten Wiuf d,*

A coalescent model of a sample of size \( n \) is derived from a birth–death process that originates at a random time in the past from a single founder individual. Over time, the descendants of the founder evolve into a population of large (infinite) size from which a sample of size \( n \) is taken. The parameters and time of the birth–death process are scaled in \( N_0 \), the size of the present-day population, while letting \( N_0 \to \infty \), similarly to how the standard Kingman coalescent process arises from the Wright–Fisher model. The model is named the Limit Birth–Death (LBD) coalescent model. Simulations from the LBD coalescent model with sample size \( n \) are computationally slow compared to standard coalescent models. Therefore, we suggest different approximations to the LBD coalescent model assuming the population size is deterministic in terms of time rather than a stochastic process. Furthermore, we introduce a hybrid LBD coalescent model, that combines the exactness of the LBD coalescent model with the speed of the approximations.

© 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

A population that originates from a single individual appears in multiple biological contexts. For example, a mutant allele might have arisen only once in the past, giving rise to a population of mutant alleles. Also, a family of homologous genes evolves from a single gene copy by gene duplication and loss (Ohta, 1989; Hurles, 2004). In addition, a population of cancer cells composing a tumor typically derive from a single precursor cell (Nowell, 1976). Furthermore, virus infections might be caused by a single particle that infects the host (Zwart et al., 2009; Zwart and Elena, 2015). Finally, species that belong to a monophyletic group share a single common ancestor in the past.

A general birth–death process appears to be an adequate framework to model such scenarios (Kendall, 1948; Kimmel and Axelrod, 2015; MacPherson et al., 2021). Since Yule's seminal paper (Yule, 1924), birth–death processes have been widely applied in biology, for example, in the context of epidemic diseases (Andersson and Britton, 2000), gene families sizes (Demuth et al., 2006), protein domain evolution (Karve et al., 2002), macroevolution (Nee, 2006), and ecology and evolution (Crawford and Suchard, 2012). In the context of evolution, typically a (small) sample is taken from a (large) population of unknown size. A main interest is often to describe the ancestry of the sampled individuals in terms of a coalescent process (Wakeley, 2008).

The birth–death process with Bernoulli sampling provides one solution to the problem of describing the sample ancestry in terms of a coalescent process (Gernhard, 2008a,b; Stadler, 2009; Lambert and Stadler, 2013; Ignatieva et al., 2020). This process is initiated at a random time in the past with a single individual that evolves forward in time according to a standard birth–death process. At the present time, a random number of individuals are chosen from the population (of unknown size), with a given probability. If the sampling probability is one, then the entire population is sampled. The coalescent process describing the ancestry of the sampled individuals under this framework is called the conditioned reconstructed process with Bernoulli sampling (CRPB) (Gernhard, 2008b). Explicit mathematical expressions can be obtained for CRPB quantities of interest, and efficient tools for Bayesian inference have been developed (Drummond et al., 2012; Drummond and Bouckaert, 2015).

However, despite its mathematical simplicity, the CRPB has its shortcomings. In many real situations, the number of sampled individuals is controlled by the researcher, but the CRPB assumes the sample size is a random variable (which might result in zero, one, ... individuals). In the CRPB, this implies that the sample size correlates positively with the population size, and hence also with the age of the founder individual (not to be confused with the most recent common ancestor (MRCA) of the sample). In other words, the age of the founder individual depends on the sample size (Gernhard, 2008b; Wiuf, 2018). If the sample size is...
large, then the founder is expected to be older than if the sample size is small, simply because the number of (randomly) sampled individuals reflects the size of the entire population.

To overcome this limitation, we propose a new approach. Our starting point is a birth–death process with a single origin at a random time in the past. At the present time, the population is assumed to be large, say, with \( N_0 \) individuals. By scaling the parameters and time in \( N_0 \) and letting \( N_0 \to \infty \), we derive a coalescent process, named the LBD coalescent model (LBD = Limit Birth–Death), of a sample of \( n \) individuals. This approach is akin to the standard (i.e., large \( N_0 \)) coalescent processes in population genetics. Conveniently, this process does not have a parameter for the (unknown) current population size, so it implies one parameter less than the original birth–death process. Furthermore, the sample size does not correlate positively with the population size as in the CRPB.

We describe this new LBD coalescent model of a sample of size \( n \), and provide means to simulate it. A similar approach has been applied in population genetics to study a population, and a sample, of rare alleles (Wiu, 2000, 2018), see also Thompson (1975). However, this strategy can be computationally slow, so we also propose different approximations inspired by the coalescent with deterministic (time-dependent) population size (Griffiths and Tavaré, 1994), approximating the stochastic population size over time with a deterministic function. One of these approximations is the coalescent model with exponential growth (Slatkin and Hudson, 1991), that has been investigated previously in connection with birth–death processes (Stadler et al., 2015; Volz and Frost, 2014). We demonstrate theoretically and by simulation that these deterministic population-size approximations behave well near the present time, but become gradually worse as the population size becomes small and stochastic fluctuations cannot be ignored. To leverage the advantages of the LBD coalescent model and the deterministic population-size approximations, we propose a hybrid approach and further an ad hoc approach that maximize the trade-off between accuracy and computational speed. Surprisingly, the ad hoc approach appears to be more efficient than the hybrid approach, both in terms of computational time and accuracy.

Simulation details are in Appendix A. Mathematical derivations are described in the Supplementary Material.

2. The standard birth–death process

Parameters. The starting point for the LBD coalescent model is a standard birth–death process (Kendall, 1948), see Fig. 1. This process has two parameters, a birth rate \( \nu \), and a death rate \( \mu \). Define

\[
\gamma = \frac{|\nu - \mu|}{\max(\nu, \mu)}, \quad \delta = |\nu - \mu|,
\]

where \( \nu > 0 \), and \( \mu > 0 \) (Wiu, 2018). We assume \( \nu \neq \mu \) for convenience, and note that \( \nu = \mu \) results in different mathematical expressions, but is otherwise not special. Hence, \((\nu, \delta) \in \mathbb{R}^2_+\). Furthermore, \((\nu, \delta)\) remains the same if the birth rate and the death rate are interchanged. Note that \( \gamma \) is a dimensionless constant, thus a characteristic of the process, while \( \delta \) depends on the chosen time scale. If the time scale is changed, \( \gamma \) remains invariant. The parameter \( \gamma \) might be interpreted as a growth rate independent of the chosen unit of time.

A population of finite size \( N_0 \). We assume the birth–death process is initiated at a random time (assuming an improper uniform prior) in the past, and evolves until the present time, when there are \( N_0 \) individuals. A reverse (coalescent) process going backwards in time, describing the ancestry of all \( N_0 \) individuals is the following, see Gernhard (2008b) and Fig. 2. Let \( T_1, T_2, \ldots, T_{N_0} \) be the coalescent times, such that \( T_1 \) is the age of the population, and \( T_k, k = 2, \ldots, N_0 \), the time while there at least \( k \) ancestors. Clearly, \( T_1 > T_2 > \cdots > T_{N_0} > 0 \), counting time backwards from the present. For convenience, we let \( T_{N_0+1} = 0 \).

The density of the increment \( T_k - T_{k+1}, k \geq 1 \), conditional on \( T_{k+1} = s \), is

\[
f_{N_0}(t|s) = (k - 1)\gamma s^\nu e^{-(\nu + \delta)t} \frac{(1 - e^{-\delta t})}{(1 + (\nu - 1)e^{-\delta t})^{k+1}},
\]

where \( \gamma_s = 1 + (\nu - 1)e^{-\delta t} \). Hence \( T_k = s + t \). If \( k = N_0 \), then \( s = 0 \) and \( \gamma_0 = \gamma \). For \( \ell = 1 \) and \( k = N_0 \), we find the density of \( T_1 \), the age of the process,

\[
f_{1,N_0}(t) = N_0 \delta \gamma e^{-(\nu - \delta) N_0 t} \frac{e^{-\delta t}(1 - e^{-\delta t} N_0 - 1)}{1 + (\nu - 1)e^{-\delta t} N_0 + 1}.
\]

For details, see Gernhard (2008b) (using a different parameterization) or (Wiu, 2018) (using the same parameterization as here).

We name the process \( T_1, T_2, \ldots, T_{N_0} \) with distribution given by (1)-(2) the birth–death coalescent process with population size \( N_0 \).

3. The LBD coalescent model with sample size \( n \)

In many biological scenarios, the population size is much bigger than the sample size, \( N_0 \gg n \). Here, we consider the situation where \( N_0 \) is large and unknown, hence it is sensible to consider the limiting model as \( N_0 \to \infty \). This approach is common in population genetics (Wakeley, 2008). Conveniently, \( N_0 \) is a parameter that is difficult to estimate reliably from empirical data and taking \( N_0 \to \infty \) reduces the number of parameters by one.

The population process described by (1)-(2) has three parameters \( N_0, \delta, \gamma \). Consider the scenario:

\[
N_0 \delta \to \Delta, \quad N_0 \gamma' \to \Gamma,
\]

where \( N_0 \delta \to \Delta \) (resp. \( N_0 \gamma' \to \Gamma \)) means we consider models such that the product of \( N_0 \) and \( \delta \) (resp. \( \gamma' \)) converges to \( \Delta \) (resp. \( \Gamma \)), while \( N_0 \to \infty \), implying \( \delta \to 0 \) (resp. \( \gamma' \to 0 \)) at the same time. Compared to a standard coalescent model, \( N_0 \) plays the role of the effective population size.

---

Fig. 1. A standard birth–death process. Left: Outcome of a birth–death process. The population size varies stochastically over time with current population size \( N_0 = 4 \). Solid lines: lineages surviving until the present time. Dashed lines: lineages dying out before present time. Right: The same figure as before but with lineages going extinct removed. One lineage is ‘ironed out’ for illustrative purposes. Compared to standard coalescent processes, the tip above the MRCA of the population represents the time from the MRCA back to the origin.
We derive the coalescent times of the whole population under the constraint (3), and use Saunders et al. (1984, Lemma 3) to derive the coalescent times of the sample in terms of those of the entire population. We name the coalescent process of the sample the LBD coalescent model (with sample size $n$), as mentioned earlier.

The limit of (1) under the constraint (3) is

$$f_{c,\infty}(v|u) = k \frac{(k-1)!}{(\ell-1)!} \frac{1 - e^{-\Delta\mu}}{1 - e^{-\Delta\mu}} \frac{(1 - e^{-\Delta\mu})^{\ell - 1}}{(1 - e^{-\Delta\mu})^{\ell - k}} \frac{1}{(1 - e^{-\Delta\mu})^{\ell - 1}},$$

and for $k = \infty$ (a population of infinite size),

$$f_{c,\infty}(u) = \Delta \Gamma^\ell \frac{1 - e^{-\Delta\mu}}{1 - e^{-\Delta\mu}} \frac{1}{(1 - e^{-\Delta\mu})^{\ell - 1}} \exp \left( - \frac{\Gamma e^{-\Delta\mu}}{1 - e^{-\Delta\mu}} \right)$$

(Wiuf, 2018). In the LBD coalescent model, time is scaled in units of $N_0$ (as $N_0 \to \infty$) as is standard in large $N_0$ coalescent approximations. Note that (5) depends on $\Gamma$, while (4) does not. We emphasize that $\Gamma$ is dimensionless and therefore invariant under rescaling of time, while $\Delta$ is not. We therefore assume $\Delta = 1$ in simulations.

Consider a sample of size $n$ taken from the population and denote the coalescent times of the sample by $T_1, \ldots, T_n$, see Fig. 2. That is, $T_k, k = 2, \ldots, n$, is the time while there are at least $k$ ancestors of the sample. Let the (stochastic) number of ancestors of the entire population be $m_k, k = 1, \ldots, n - 1$, the first time there are $k$ ancestors of the sample. Hence, $T_k = T_{m_k+1}$. When there are at least $k$ ancestors of the sample, there are at least $m_k + 1$ ancestors of the population, see Fig. 2.

According to Saunders et al. (1984, Lemma 3), the distribution function of $m_k$, conditional on $m_k = k, k = 2, \ldots, n - 1$, is

$$F_k(x|m_k) = \frac{(m_k - k)! (m_k + k - 1)!}{m_k!(m_k - 1)!} \frac{x(x - 1) \ldots (x - (k - 2))}{(x + k)(x + k - 1) \ldots (x + 2)}.$$

The lineages that coalesce are chosen uniformly among the ancestral lineages, similarly to other neutral coalescent processes (Kingman, 1982). Thus, to draw a genealogy of a sample of size $n$ from the LBD coalescent model, we first draw the numbers $m_k, k = 1, \ldots, n - 1$, from (6)–(7), and then draw the times $T_k = T_{m_k+1}, k = 2, \ldots, n$ from (4)–(5). Simulation details are provided in Appendix A.

The limit expressions (4)–(5) are in good agreement with those of the birth–death coalescent process with population size $N_0$ (Kingman, 1982), even for relatively small $N_0$, such as $N_0 = 100$. This is illustrated in Figure S1 in the Supplementary Material, where the time of the origin, $T_1$, and the time of the MRCA of the population, $T_2$, respectively, are compared for different values of $N_0$, including the limit $N_0 \to \infty$. The approximation accuracy depends on $\Gamma$, but not $\Delta$, as $\Delta$ just scales time.

We end this section by pointing out related work in the literature. In Stadler (2009), an expression is derived from (1)–(2) for the density of the coalescent time of a sample of size $n = 2$, with and without taking the limit as $N_0 \to \infty$. In Harris et al. (2019), continuous-time Galton–Watson processes are discussed. The coalescent times of a sample of fixed size, say $n$, taken from a population of unknown size, are derived, assuming the time of the origin $T_1$ is a fixed parameter, unlike in our case where we put a prior distribution on $T_1$. Some general results are derived in the critical case $\nu = \mu$ (which is not covered in this paper), and in the scaling limit as $T_1 \to \infty$, assuming the mean number of offspring is $1 + \frac{\mu}{\Gamma}, m \in \mathbb{R}$. The mean number of offspring converges to one with increasing $T_1$. In Grosjean and Huillet (2018), the transition structure and coalescence times of a sample of fixed size are discussed for general Bienaymé–Galton–Watson processes, assuming the origin is a fixed parameter. A different scaling limit is considered in Ignatieva et al. (2020) in the context of the CRPB with vanishing sampling probability, $\psi \to 0$. As a consequence, the current population size increases towards infinity in the limit. Exact expressions are obtained for distributions of interest. This approach, though mathematically very appealing, has the same problem as the CRPB, namely, that the age of the founder individual depends on the sample size. Finally, discrete-time Galton–Watson processes are analysed in Burden and Soewongsono (2019) in the diffusion limit as $\nu \to 1$ and $T_1 \to \infty$ (in a particular way), while the number of founder individuals becomes infinitely large.

4. Approximations to the LBD coalescent model

In the LBD coalescent model we have to simulate the number of ancestors of the entire population at each coalescent event of the sample, which is computationally intensive. To simplify simulations and speed up computations, we next propose three different approximations to the LBD coalescent model that assume a coalescent process with time-varying deterministic population size (Griffiths and Tavaré, 1994). This avoids the hassle of simulating the number of ancestors of the entire population at each coalescent event of the sample.

We will argue heuristically for the approximations. Let $N(t)$ denote the stochastic population size at time $t$ in the past, and let $N(0) = N_0$. For the moment, we assume time is unscaled. Assume there are $2 \leq k \leq n$ ancestors of a sample of present size $n$. If a coalescent event happens at time $t$ in the past (which is a birth event in the forward direction of time), then the probability that the coalescent event is between two of the $k$ ancestors, is

$$\frac{k(k - 1)}{(N(t) + 1)N(t)}$$
(Saunders et al., 1984) (there are \(N(t) + 1\) individuals in the population just prior to the coalescent event). By conditioning on the population size process, \(N = (N(t), t \geq 0)\), the probability that a coalescent event has not happened before time \(T_{k+1} = s\) and time \(t > s\), is

\[
P_{\Lambda}(T_k > t|T_{k+1} = s) = \left( \prod_{i=1}^{k+1} \left( 1 - \frac{k(k-1)}{(N(s_i + 1) + 1)N(s_i)} \right) \right) \approx \exp \left\{ -\sum_{i=1}^{k+1} \frac{k(k-1)}{(N(s_i + 1) + 1)N(s_i)} \right\} = \exp \left\{ -\frac{k}{2} \int_s^t \frac{2\nu}{N(z) + 1} dz \right\}.
\]

assuming the population size is large. Here \(B_{s,t}\) denotes the number of birth events between time \(s\) and \(t\), \(s_i\) denotes the time of the \(i\)th birth event, and the number of birth events within a small time interval is of order \(N(z)\) in probability.

By taking the expectation over the population size process, we obtain

\[
P_{\mu}(T_k > t|T_{k+1} = s) \approx E_{N_0} \left[ \exp \left\{ -\frac{k}{2} \int_s^t \frac{2\nu}{N(z) + 1} dz \right\} \right].
\]

where \(E_{N_0}\) denotes the expectation over all possible outcomes of the population size process, assuming \(N(0) = N_0\). Later, we also condition on the age of the origin, in which case the expectation in (8) is over all outcomes of the population size process with a given age. By Jensen's inequality,

\[
P_{\mu}(T_k > t|T_{k+1} = s) \geq \exp \left\{ -\frac{k}{2} \int_s^t E_{N_0} \left( \frac{2\nu}{N(z) + 1} \right) dz \right\},
\]

where \(\geq\) means approximately greater than. In (9), we have (approximate) equality if the population sizes are large and the stochastic fluctuations comparably small (Sjodin et al., 2005; Krone and Kai, 2003). In fact, we later show this is the case for small (population scaled) times, see (18), whereas it fails for large times. Also by Jensen's inequality,

\[
\exp \left\{ -\frac{k}{2} \int_s^t E_{N_0} \left( \frac{2\nu}{N(z) + 1} \right) dz \right\} \lesssim \exp \left\{ -\frac{k}{2} \int_s^t \frac{2\nu}{N_0(N(z) + 1)} dz \right\},
\]

with the same amendments about equality as before. The right hand side in (10) might in some cases be closer to the probability \(P_{\mu}(T_k > t|T_{k+1} = s)\) than the left hand side. In fact, as the function \(x \mapsto e^{-x^2/2}\) is convex for \(0 < x < \frac{1}{2}\), the right hand side of (10) will be larger than \(P_{\mu}(T_k > t|T_{k+1} = s)\) whenever the exponent of the exponential is smaller than \(\frac{1}{2}\). We note that the factor \(2\nu\) is consistent with Volz et al. (2009).

If the population size is large and stochastic effects comparatively small, then equality is expected above (Sjodin et al., 2005; Krone and Kai, 2003). In that case, by scaling time in units of \(N_0\), assuming \(N_0\delta \rightarrow \Delta\) and \(N_0\gamma \rightarrow \Gamma\), as in (3), the right hand sides of (9) and (10) give rise to time-varying population size coalescent processes (Griffiths and Tavare, 1994). In a time-varying population size coalescent process, the random variable \(T_k\), conditional on \(T_{k+1} = u\), has density and distribution function, respectively, given by

\[
g_k(v|u) = \left( \frac{k}{2} \right) \lambda(v) \exp \left\{ -\frac{k}{2}(\Lambda(v) - \Lambda(u)) \right\},
\]

for \(k = 2, \ldots, n\), where \(T_{k+1} = 0\) for convenience, \(\lambda(u)\) is the coalescent intensity at time \(u \geq 0\), and \(\Lambda(v) = \int_u^v \lambda(u) du\), the accumulated intensity (Griffiths and Tavare, 1994). Taking \(\lambda(u) = u\), we retrieve the standard Kingman coalescent. See Appendix A on how to simulate from a time-varying population size coalescent process (11).

In our case, the coalescent intensities are derived from the exponents of the right hand sides of (9)–(10);

\[
a) \quad E_{N_0} \left( \frac{2\nu}{N(z) + 1} \right), \quad b) \quad E_{N_0} \left( \frac{2\nu}{N_0(N(z) + 1)} \right).
\]

The limits are derived in the Supplementary Material, Sections 2 and 3, and discussed below. Before providing the details, we note that

\[
v = \frac{N_0\delta}{N_0\gamma} \quad \text{if} \quad v \geq \mu,
\]

\[
v = \mu - \delta = \frac{N_0\delta}{N_0\gamma} - \delta \quad \text{if} \quad v < \mu,
\]

hence, in either case,

\[v \rightarrow \frac{\Delta}{\Gamma} \quad \text{as} \quad N_0 \rightarrow \infty.\]

Approximation 1: The \(M_a\) coalescent model, based on (14)(a).

To take some of the stochasticity in the population size process into account, we further condition on the age of the process. The coalescent intensity at time \(u\) given the age \(T_1\) is the large \(N_0\) limit of (14)(a),

\[
\lambda_a(u, T_1) = \lim_{N_0 \rightarrow \infty} E_{N_0} \left( \frac{2\nu}{M_0(N_0(N_0u) + 1)} \right) T_1
\]

\[
= \frac{2\Delta e^{\Delta u} (1 - e^{-\Delta T_1})^2}{\Gamma} \left( 1 - e^{-\frac{\Gamma}{1 - e^{-\Delta T_1}}} \right),
\]

for \(0 \leq u \leq T_1\), see the Supplementary Material (Sections 2 and 3) for details.

As \(\lambda_a(u, T_1) \rightarrow \infty\) as \(u \rightarrow T_1\), then it is guaranteed that a MRCA of the sample is found eventually (Griffiths and Tavare, 1994). This accumulated intensity \(\lambda_a(u, T_1) = \int_0^u \lambda_a(s, T_1) ds\) cannot be given in explicit form, hence we apply numerical approximations to compute it.

Approximation 2: The \(M_b\) coalescent model, based on (13)(b). For the second approximation, we also condition on the age of the process. The average relative population size at time \(u\) given the age \(T_1\) is

\[
h(u, T_1) = \lim_{N_0 \rightarrow \infty} E_{N_0} (M_0(N_0u) | T_1)
\]

\[
= \frac{e^{-\Delta u} (1 - e^{-\Delta T_1 - u})^2}{(1 - e^{-\Delta T_1})^2} + \frac{2}{\Gamma} \left( 1 - e^{-\Delta T_1 - u} \right) \left( 1 - e^{-\Delta u} \right).
\]
for $0 \leq u \leq T_1$, see the Supplementary Material (Sections 2 and 3) for details of the derivation. The relative population size at time $T_1$ is $h(T_1, T_1) = 0$, which is the limit of $1/N_0$ as $N_0 \to \infty$. The coalescent intensity at time $u$ given the age $T_1$ is the large $N_0$ limit of (13),

$$\lambda_2(u, T_1) = \frac{2\Delta}{T} \frac{1}{h(u, T_1)} = \Delta e^{\Delta u} \left( \frac{1}{e^{\Delta T_1} - e^{\Delta u}} + \frac{\alpha_z}{\alpha_2 e^{\Delta u} + b_2} \right),$$

where

$$\alpha_z = 2 \left( 1 - e^{-\Delta T_1} \right) - e^{-\Delta T_1}, \quad b_2 = \frac{1}{2} \left( 1 - e^{-\Delta T_1} \right).$$

In this case, the accumulated intensity can be found explicitly,

$$\Lambda_2(s, T_1) = \int_0^s \lambda_2(u, T_1) du = \log \left( \frac{\alpha_2 e^{\Delta u} + b_2}{1 - e^{-\Delta(T - s)}} \right).$$

As $h(T_1, T_1) = 0$, then $\lambda_2(u, T_1) \to \infty$ as $u \to T_1$, and it is guaranteed that a MRCA of the sample is found before time $T_1$ (Griffiths and Tavaré, 1994).

**Approximation 3: The $M_2$ coalescent model.** We suggest one further heuristic approximation, that allows the coalescent intensity to be increased or decreased, relatively to $M_2$ and $M_3$. It provides one way to repair the shortcomings of the $M_2$ and $M_3$ coalescent models (to be discussed later).

The coalescent intensity, $\lambda_2(u, T_1)$, for $M_2$ might alternatively be written as

$$\lambda_2(u, T_1) = \frac{2\Delta}{T} \frac{1}{h(u, T_1)} \left( 1 + 2 \left( 1 - e^{-\Delta T_1} \right) \left( e^{\Delta u} - 1 \right) \right)^{-1}.$$

This suggests a suit of approximations, with intensities given by

$$\lambda_k(u, T_1) = \frac{2\Delta}{T} \frac{1}{h(u, T_1)} \left( 1 + K \left( 1 - e^{-\Delta T_1} \right) \left( e^{\Delta u} - 1 \right) \right)^{-1},$$

$K > 0$. For small $u$, the intensities of $M_K$ and $M_2$, respectively, $M_K$ and $M_2$, agree,

$$\lambda_k(u, T_1) \approx 1 \quad \lambda_k(u, T_1) \approx 1, \quad u \approx 0,$$

while for large $u$, we find

$$\lambda_k(u, T_1) \approx \frac{2}{K} \quad \lambda_k(u, T_1) \approx \frac{1}{K}, \quad u \approx T_1.$$

Thus, by choosing $K$, one might speed up or slow down the coalescent intensity by a constant factor for large $u$, compared to $M_2$ and $M_3$, respectively. In any case, $\lambda_k(u, T_1) \geq \lambda_2(u, T_1)$ for $K \leq 2$. See Fig. 3 for an illustration.

For the accumulated intensity of $M_K$, we find

$$\Lambda_k(s, T_1) = \frac{2}{K} \log \left( \frac{\alpha_2 e^{\Delta u} + b_k}{1 - e^{-\Delta(T - s)}} \right),$$

with

$$\alpha_k = \frac{K}{2} (1 - e^{-\Delta(T_1)}) - e^{-\Delta(T_1)}, \quad b_k = 1 - \frac{K}{2} (1 - e^{-\Delta(T_1)}).$$

### 5. The hybrid LBD coalescent model

The $M_2$ coalescent model is computationally inefficient, while $M_2$ might be unsatisfactory compared to the LBD coalescent model (to be discussed later). Therefore, it might be interesting to explore a hybrid approach between the LBD coalescent model and the $M_2$ coalescent model. To do this, for a given sample size $n$, we first simulate the coalescent times according to the $M_2$ coalescent model until there are $k_0$ ancestors of the sample at time $T_{k_0+1}$, for a fixed $k_0 \leq n$. Then we continue with the LBD coalescent model from time $T_{k_0+1}$. For this, we need to know the number of ancestors $A_0$ of the whole population in the $M_2$ coalescent model at time $T_{k_0+1}$.

The number $A_0$ is estimated from the equation

$$\frac{2}{A_0} = A_2(\sim T_{k_0+1}, T_1),$$

noting that $A_2(\cdot, T_1)$ provides a correspondence between the time in the $M_2$ coalescent model and the time in the standard Kingman coalescent model with constant population size. Such relationship is generally used to simulate times from a time-varying population size coalescent model (Griffiths and Tavaré, 1994). See the paragraph The $M_2$ and $M_k$ coalescent models, $K > 0$, in the Appendix A for details. The value $2/A_0$ in (15) is the expectation of $T_{k_0+1}$, the time until there are $A_0$ ancestors to the whole population in the standard Kingman coalescent model. If $A_0$ is large (enough), the time $T_{k_0+1}$ has negligible standard deviation compared to its expectation, making $2/A_0$ an accurate estimate of $U_{k_0+1}$, hence also of the corresponding time $T_{k_0+1}$ in the LBD coalescent model, see the Supplementary Material (Section 5) for details of the calculation. From time $T_{k_0+1}$ onwards the coalescent times are sampled according to the LBD coalescent model, letting $m_{k_0} = A_0$ initially.

In the LBD coalescent model, the number of ancestors $m_k$ of the entire population the first time there are $k$ ancestors of the sample is time-consuming to simulate. Furthermore, the expectation of $m_{n-1}$ is infinite, while it is $nk/(n - k - 1)$, for $k = 1, \ldots, n-2$, see the Supplementary Material (Section 1) for a demonstration. This suggests taking $k_0$ to be small. However, one cannot choose $k_0$ too small as this increases the probability that $A_0$ in (15) is smaller than $k_0$, which leads to an impossibility.

A similar “hybrid” approach is taken in Bhaskara et al. (2014) in a study of the accuracy of the standard Kingman coalescent model for small population sizes.

### 6. Simulations

**Implementation.** We implemented functions in R (https://cran.r-project.org/) to simulate from the LBD coalescent model and the different approximations, including the hybrid LBD coalescent model. The R code is available at https://github.com/crespofabian8012.

We investigated the LBD coalescent model and the various approximations for two different sample sizes ($n = 10,100$) with fixed $\Delta = 1$, and varying values of $\Gamma$. For all choices of parameters, $10^4$ simulations of ancestry were obtained (except for assessment of the running time). The effect of $\Delta$ is a time scale transformation, hence there is no point in varying $\Delta$. For the hybrid approach we took $k_0$ to be 50% or 25% of $n$, that is, $k_0 = [n/2]$ and $k_0 = [n/4]$.

The effect of $\Gamma$ in the LBD coalescent model. Table 1 shows the effect of $\Gamma$ on the genealogy relating a sample of individuals. Both the relationship between the age and the time of the MRCA as well as the relationship between branch lengths depend on $\Gamma$. Eventually, for large $\Gamma$, the tree becomes star shaped, and the time gap between the MRCA and the origin shrinks to zero. Fig. 4 shows single realizations of genealogies.

**Computational time.** For the LBD coalescent model and the hybrid approach, the main determinant of computational speed is the simulation of the number of ancestors $m_k$. As the distribution of this quantity is independent of $\Gamma$, the overall computational time is independent of $\Gamma$. The same is the case for the approximations $M_2$, $K > 0$, while for $M_k$, computational time might depend on $\Gamma$ through the concrete numerical procedure applied, though this does not seem to be an issue with our implementation.
Fig. 3. Coalescent intensities. Shown is the coalescent intensity $\lambda_*(u, T_1)$ (black), $\lambda_{BS}(u, T_1)$ (green), $\lambda_1(u, T_1)$ (red), and $\lambda_2(u, T_1)$ (blue) for $\Gamma = 0.1$ (top row) and $\Gamma = 10$ (bottom row), and two random draws of $T_1$ for each $\Gamma$ (and $\Delta = 1$), according to (5). The coalescent intensities are not necessarily increasing functions over time, in particular, this might not be the case for small $\Gamma$ and large $T_1$ (which is unlikely for small $\Gamma$). This is even so for the intensities of $M_1$ and $M_2$. The intensity of $M_K$ decreases with increasing $K$.

Fig. 4. Examples of genealogies and lineages-through-time plots. One example for each of $\Gamma = 0.001, 0.1, 10, 1000$, and $n = 10$. Top panel: Genealogies. Bottom panel: The number of lineages in the genealogy (normalized to one) against normalized time. As $\Gamma$ increases the genealogy becomes more and more star-shaped.
The reduction (or increase) in computational time compared to the LBD coalescent model is shown in Table 2. In particular for the hybrid LBD coalescent model with \( n = 10 \), it happens that \( k_0 > A_0 \), which causes an impossibility (the number of ancestors of the sample cannot be larger than the number of ancestors of the population). Since, we need to exclude these cases, this causes a bias towards genealogies with large \( A_0 \).

The coalescent intensities and the relative population size. Examples of coalescent intensities for \( M_2 \), \( M_2 \) and \( M_2 \) are given in Fig. 3. It appears the coalescent intensities are not necessarily monotonically increasing, and might first decrease, then increase. For \( M_2 \), this implies the opposite pattern occurs for the relative population size \( h(u, T_1) \). However, this seemingly paradoxical result disappears when averaging over the distribution of \( T_1 \), given in (5),

\[
H(u) = \int_u^{\infty} h(u, t) f_1(t) dt = e^{-\Delta u}, \quad u \geq 0,
\]

(16)

see the Supplementary Material (Section 4) for an evaluation of the integral. The function \( H(u) \) is strictly decreasing, and does not depend on \( \Gamma \) in contrast to \( h(u, T_1) \).

One might construct yet another time-varying population size coalescent process from (16) with intensity coinciding with the intensity of the coalescent with exponential growth (Slatkin and Hudson, 1991),

\[
\lambda_2(u) = \frac{2\Delta}{\Gamma} \frac{1}{H(u)} = \frac{2\Delta}{\Gamma} e^{\Delta u}
\]

(17)

(where the factor \( \frac{2\Delta}{\Gamma} \) in front is a time scale adjustment, similar to Volz et al. (2009)). This model performs worse than the \( M_2 \) coalescent model (see Figure S2 in the Supplementary Material) and is not pursued further here. The model has the additional drawback that the age \( T_1 \) is not directly built into it. In contrast, a simulated coalescent history of a sample from one of the coalescent models described in this paper contains the coalescent times of the sample in addition to the age of the origin.

Accuracy of the \( M_2 \) and \( M_2 \) coalescent models. Fig. 5 shows the ratio of the median coalescent times for \( M_2 \) and \( M_2 \) coalescent models, compared to the LBD coalescent model. Both approximations are quite accurate near the present, but neither approximation is able to reproduce the distributions of the LBD coalescent times near the origin, with \( M_2 \) underestimating and \( M_2 \) overestimating the corresponding values. These findings are in agreement with the bounds (9) and (10), and might be attributed to the fact that large \( N_0 \) coalescent approximations are more accurate when the population size is (in fact) large. Also, when the population size is small, stochastic fluctuations are not negligible and increase the variance of the coalescent times (Sjödin et al., 2005; Krone and Kai, 2003). As the LBD coalescent model eventually ends with a single individual, large \( N_0 \) approximations are expected to fail eventually for both \( M_2 \) and \( M_2 \). We note that the approximations seem to be more accurate for large values of \( \Gamma \) and \( n \).
Average running times. Shown is average running times for the LBD coalescent model in milliseconds, and the average running times in % of the LBD mean running time for the other coalescent models, for \( n^{'} = 10 \). For the hybrid LBD coalescent model (\( H \)), it is stated how often the condition \( k_0 > 0 \) occurred. These cases are excluded. Similar running times are obtained for other values of \( n^{'} ; n \): sample size, sim: number of simulations. The simulations were executed on a MacBook Pro Intel Core i5, 2.3 GHz, 8 GB RAM 2133 MHz LPDDR3, using a single core.

Table 2

<table>
<thead>
<tr>
<th>( n )</th>
<th>( n^{'} )</th>
<th>( \text{LBD} )</th>
<th>( M_r )</th>
<th>( M_{0.8} )</th>
<th>( M_2 )</th>
<th>( k_0 = \lfloor \frac{\sqrt{t}}{2} \rfloor )</th>
<th>( k_0 = \lfloor \frac{\sqrt{t}}{1} \rfloor )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1000</td>
<td>3.6</td>
<td>19%</td>
<td>19%</td>
<td>44%</td>
<td>12</td>
<td>27% 9</td>
</tr>
<tr>
<td>100</td>
<td>1000</td>
<td>180</td>
<td>7.7%</td>
<td>7.7%</td>
<td>33%</td>
<td>16</td>
<td>18% 0</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>3100</td>
<td>4.2%</td>
<td>4.2%</td>
<td>32%</td>
<td>0</td>
<td>14% 0</td>
</tr>
</tbody>
</table>

Fig. 6. Coefficient of variation. Shown is four examples for \( n^{'} = 0.1 \) (red, orange) and \( n^{'} = 10 \) (blue, green) with \( t^* \) drawn randomly from the distribution (5). Time is scaled in units of \( t^* \) for comparison, hence it runs from 0 to 1. The coefficient of variation always approaches \( 1/\sqrt{2} \approx 0.70 \) as the scaled time approaches one, but the speed by which it does so depends on \( n^{'} \) as well as on the concrete realization of \( t^* \).

To illustrate these observations further, we found the coefficient of variation as \( N_0 \to \infty \),

\[
CV(u, t^*_i) = \lim_{N_0 \to \infty} \frac{\sqrt{\text{Var}_N(N_{0}|u)|t^*_i)}}{E_N(N_{0}|u)|t^*_i}}
\]

(see the Supplementary Material, Section 3, for the derivation). For \( u \approx t^*_i \), \( CV(u, t^*_i) \approx 1/\sqrt{2} \approx 0.7 \), hence not only is the population size small, it is also far from being constant, see Fig. 6.

Accuracy of the \( M_k \) coalescent model. Fig. 7 shows the performance of the \( M_{0.8} \) coalescent model, where \( K = 0.8 \) minimizes the squared deviation between median coalescent times of the LBD coalescent model and those of the \( M_k \) coalescent model (see Figures S3 and S4 in the Supplementary Material), for the three values \( K = 0.001, 0.1, 10 \). The same value of \( K \) was found for \( n = 10 \) and \( n = 100 \) (with the chosen precision). The median coalescent times are generally in very good agreement with the LBD coalescent model (note the y-scale in Fig. 7).

In addition, we calculated the correlation coefficients for pairs of coalescent times for the \( M_{0.8} \) coalescent model and the LBD coalescent model, and found that these generally are in good agreement with each other, see Figures S5–S7 in the Supplementary Material for \( n = 100 \). This agreement is best near the present time and gradually worsen deeper into the genealogy.

We chose the same \( K \) for all \( n \) and \( n^{'} \). Simulations indicate the best \( K \) depends to some degree on \( n \) and \( n^{'} \), though the dependence seems to be weak for the values investigated. This is also seen in Figures S5–S7 in the Supplementary Material.

The hybrid LBD coalescent model. The \( M_k \) coalescent model, for any \( K > 0 \), is generally much faster to simulate from than the LBD coalescent model (Table 2). This makes the hybrid LBD coalescent model speed-efficient as well, though the speed reduction depends on the chosen \( k_0 \).

Fig. 8 shows that the median coalescent times of the hybrid LBD coalescent model with \( K = 2 \) fits well to the median coalescent times of the LBD coalescent model (only \( n = 100 \) is shown). For the correlation coefficients, we find similar agreement to that of the \( M_{0.8} \) coalescent model, see Figures S5–S7 in the Supplementary Material for \( n = 100 \). However, surprisingly, the correlation coefficients appear to be (slightly) less accurate than those of \( M_{0.8} \). One possible explanation for this might be that the distribution of \( m_k \) (the number of ancestors of the population when the hybrid approach changes from \( M_k \) to the LBD coalescent model) differs from that of the pure LBD coalescent model. For the hybrid LBD coalescent model, the simulated mean (variance) of \( m_k \) is 13.42 (59.66) for \( n = 10 \) and \( k_0 = 5 \), and 102.86 (255.62) for \( n = 100 \) and \( k_0 = 50 \), compared to the theoretically determined mean (variance) of \( m_k \) in the LBD coalescent model, 12.5 (56.25) for \( n = 10 \) and 102.04 (219.05) for \( n = 100 \) (see Supplementary Material, Section 1, for a derivation of the theoretical mean and variance of \( m_k \)).

7. Discussion

We have proposed a coalescent framework based on a birth–death process under the constraint of a large (infinite) present-day population size, that we named the LBD coalescent model. To simulate from this model with sample size \( n \), one needs to simulate the number of ancestors of the entire population at each coalescent event of the sample. This appears difficult to do efficiently. To overcome this, we suggest efficient approximations in terms of accuracy and computational time. However, as the population size is stochastic and eventually decreases towards one, these approximations eventually fail when time is near the age of the population. To repair this deficiency, we proposed alternative models and a hybrid model approach, that combines the LBD coalescent framework with the \( M_2 \) coalescent model. We showed using simulation that both the \( M_k \) (for \( K = 0.8 \)) coalescent model and the hybrid LBD coalescent model are efficient and accurate, while the \( M_2 \) and the \( M_k \) coalescent models are less accurate.

When using a time-varying population size approach, the stochastic nature of the birth–death process is partly ignored. In the present case, this problem is two-fold. On one hand, the population size varies stochastically. On the other hand, the population size decreases eventually to one. Both of these problems run counter to obtaining accurate time-varying population size coalescent models, as also addressed previously (Stadler et al., 2015; Krone and Kai, 2003; Griffiths and Tavaré, 1994).

Our approach shows one important difference to the CRPB in that the coalescent rates of the sample are quadratic in the number of lineages instead of linear. This difference is caused by the stochastic nature of the sample size in the CRPB (Gernhard, 2008b; Wiuf, 2018), where each sampled individual contributes a geometric random number to the entire population size (i.e., \( \sum_{i=1}^{n} X_i \), where \( X_i \) are iid geometric random variables); hence the entire population size scales linearly in the number of sampled individuals and the coalescent rates become linear in the sample size. In contrast, the LBD coalescent model assumes the
current population size is fixed, though large ($N_0 \to \infty$). Hence the stochasticity in the current population size is controlled.

In Stadler et al. (2015), a birth–death process is compared to a coalescent model with exponentially growing population size (for a sample of size $n = 2$), determined by the expectation of the population size of the birth–death process running forward in time. This approach is different from our approach, though it results in the same coalescent intensity as in (17). It assumes a fixed age and a stochastically varying population size, including a stochastic current population size. One might see our approach as an attempt to control this variation by conditioning on the age of the origin (while keeping it stochastic) and fixing the present day population size, while letting it be large.

One drawback of the coalescent models $M_K$, $K > 0$, is that the best $K$ depends to some extent on $n$ and $\Gamma$. This implies one might have to readjust $K$, for example, if $\Gamma$ is updated stochastically in a Bayesian inference scheme. In contrast, time-varying population size coalescent approximations, such as $M_K$, $K > 0$, facilitates using much of the inference machinery developed over the past twenty (or so) years for coalescent processes, including importance sampling (Griffiths and Tavaré, 1994), ABC (Cornuet et al., 2014) and more advanced Bayesian approaches (Drummond et al., 2012; Drummond and Bouckaert, 2015). Moreover, heterochronous sampling, which for example is relevant for modelling virus data, is relatively simple to handle in time-varying population size coalescent models (Drummond et al., 2005). It is considerably harder to handle heterochronous sampling in the LBD coalescent model, where the entire population size at any time in the past is stochastic and finite (Wiuf, 2018). Finiteness implies that individuals included, say at time $u > 0$ in the past, might coincide with ancestors of samples included prior to time $u$, or that more individuals should be included than there are individuals in the population at that time. These aspects naturally put constraints on sampling-based inference procedures for the LBD coalescent model.

It is noteworthy, that one cannot infer the parameter $\Delta$ and the mutation rate, say $\theta$, separately (assuming data are DNA sequences), but only jointly as a compound parameter $\theta/\Delta$. It follows by rescaling time by $\Delta$, hence the mutation rate in the new unit is $\theta/\Delta$. In contrast, the parameter $\Gamma$, which is a scaled growth rate independent of the time unit, might be inferred. This is the case both for the LBD coalescent model and its approximations.

The LBD coalescent model is a relatively simple population process. However, our work demonstrates the difficulties in providing fast and reliable approximations in terms of time-varying population size coalescent models.

CRediT authorship contribution statement

Fausto F. Crespo: Software, Validation, Formal analysis. David Posada: Conceptualization, Resources, Writing - review & editing, Funding acquisition. Carsten Wiuf: Conceptualization, Methodology, Validation, Formal analysis, Writing - original draft, Funding acquisition.

Acknowledgements

This work is supported by the Independent Research Fund Denmark (CW), ITN-CONTRA EU grant H2020 MSCA-ITN-2017-766030 (FC, DP), and ERC-617457-PHYLOCANCER (DP). We thank two anonymous reviewers for their valuable comments and suggestions, that improved the presentation of the manuscript.
f

a unique solution exists. (that is, assume $x$ and output solve $F$.)

To simulate from $F$, let $\mu = \log(1 - x)$, we use the approach described above for the LBD coalescent model. $k = 2, \ldots, n$, intensity $\lambda(u)$, and accumulated intensity $\Lambda(v) = \int_0^v \lambda(u) du$, we follow Griffiths and Tavaré (1994):

(a) Simulate coalescent times from the standard Kingman coalescent model, $V_k \sim \text{Exp}(k(k - 1)/2)$, $k = 2, \ldots, n$.

(b) Define $U_k = \sum_{i=k}^n V_i$, $k = 2, \ldots, n$.

(c) Solve $U_k = \Lambda(T_k)$ for $T_k$, $k = 2, \ldots, n$.

Then, $T_k$, $k = 2, \ldots, n$, have distribution in accordance with the variable population size coalescent process.

For (c), one needs to solve $u = \Lambda(t)$, that is, to determine the inverse of $\Lambda$ (by definition $\Lambda(t)$ is increasing in $t \geq 0$ such that the inverse exists). In model $M_\alpha$, the inverse of $\Lambda_k(\cdot, T_1)$ is

$$s = \Lambda_k^{-1}(t, T_1) = \frac{1}{\Delta} \log \left( \frac{x - b_k}{x e^{-\Delta t_1} + a_k} \right).$$

for $K > 0$, with $a_k, b_k$ as defined previously, and $x = e^{\frac{K}{\Delta}}$.

In the $M_\alpha$ coalescent model, there is not an exact solution, not even a closed expression exists for $\Lambda_k(t, T_1)$, hence the solution must be found by approximate/numerical means.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.tpb.2021.09.003.
References


