PNAS RESEARCH ARTICLE | EVOLUTION

A class of identifiable phylogenetic birth–death models

Brandon Legried[a] and Jonathan Terhorst[a,1]

Edited by Marcus Feldman, Stanford University, Stanford, CA; received October 25, 2021; accepted July 22, 2022

In a striking result, Louca and Pennell [S. Louca, M. W. Pennell, *Nature* 580, 502–505 (2020)] recently proved that a large class of phylogenetic birth–death models is statistically unidentifiable from lineage-through-time (LTT) data: Any pair of sufficiently smooth birth and death rate functions is "congruent" to an infinite collection of other rate functions, all of which have the same likelihood for any LTT vector of any dimension. As Louca and Pennell argue, this fact has distressing implications for the thousands of studies that have utilized birth–death models to study evolution. In this paper, we qualify their finding by proving that an alternative and widely used class of birth–death models is indeed identifiable. Specifically, we show that piecewise constant birth–death models can, in principle, be consistently estimated and distinguished from one another, given a sufficiently large extant timetree and some knowledge of the present-day population. Subject to mild regularity conditions, we further show that any unidentifiable birth–death model class can be arbitrarily closely approximated by a class of identifiable models. The sampling requirements needed for our results to hold are explicit and are expected to be satisfied in many contexts such as the phylodynamic analysis of a global pandemic.

identifiability | birth–death models | phylogenetics | phylodynamics

Significance

Thousands of publications have utilized phylogenetic birth–death models to make inferences about fundamental evolutionary processes like speciation, extinction, and the spread of pathogens. A recent study has called these findings into question, by showing that a variety of different evolutionary hypotheses are consistent with any given dataset and cannot be distinguished from one another regardless of how many additional data are collected. Here, we qualify this grave result by proving that a parsimonious subset of the evolutionary hypothesis space can be inferred from data.

The birth–death process (1, 2) is a classic model of population growth. Recently, it has also been used to study speciation and extinction (3–6) and also the evolution of pathogens (7). Data-driven inquiry in these fields is inherently challenging, because the majority of species and pathogens that ever lived have left us with no record of their existence. Thus, we can only make inferences about evolution on the basis of a biased sample of the species or lineages that happened to survive to the present day (6, 8). Interest in the birth–death process arises in part from the fact that it provides a principled way of correcting this bias (9, 10).

Realizations of the birth–death process can be viewed from a phylogenetic perspective as rooted trees, where each leaf node represents a species that survived until the present, internal nodes are unobserved species, and edges represent lines of descent. The shape of the tree is governed by two nonnegative functions that describe, at any given time t before the present, the per-capita rates of birth and death. As noted above, a distinguishing feature of this model is that lineages that died out before the present are not reflected in the resulting tree. Given birth and death rates, as well as a third parameter known as the sampling fraction, we refer to the resulting distribution over random trees as a phylogenetic birth–death (henceforth, BD) model. (A precise definition is given in the next section.) The BD model implies a distribution over observed evolutionary data, and given such data, we can use statistical estimation to make inferences about the model parameters.

BD models have been utilized in thousands of published studies (11–13), despite possessing known and somewhat troubling limitations. Stadler (14) showed there exist different birth–death models that have the same likelihood in terms of observable data. In statistical terms, this implies that the BD model is unidentifiable without further assumptions. The models considered by Stadler are highly parsimonious, consisting of constant birth and death rates that do not change over time. The problem is made even more challenging if the rates are time varying (15).

Very recently, Louca and Pennell (16) (cited hereafter as LP) proved that the situation is actually much worse than was previously realized: For any reasonably smooth birth and death rate functions, there are infinitely many other such functions that result in the same distribution over phylogenetic trees. Although each of these functions represents a qualitatively different evolutionary scenario, LP's result shows that it is impossible to tell which of them produced a given dataset, even if the data were infinite. In light of the huge number of times that this model has appeared in the literature, this finding is highly worrisome.

Consistent estimation is impossible in an unidentifiable statistical model, so when faced with one, there are two ways forward: 1) Use a different model, or 2) impose additional regularity conditions on the parameter space to restore identifiability. For the BD model,

Author affiliations: [a]Department of Statistics, University of Michigan, Ann Arbor, MI 48109

Author contributions: B.L. and J.T. designed research, performed research, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: jonth@umich.edu.

Published August 22, 2022.

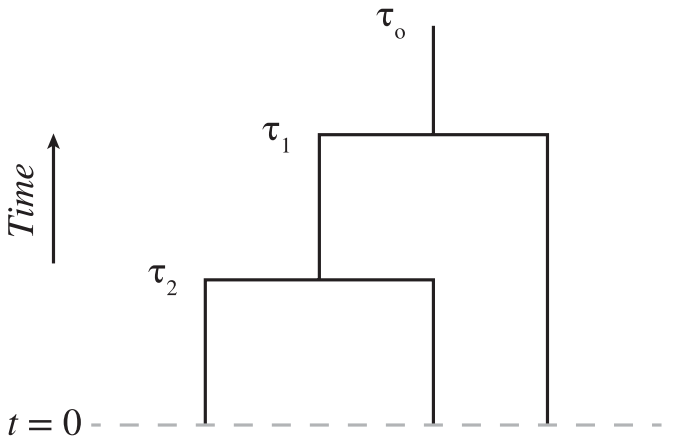

Fig. 1. An extant timetree on $n + 1 = 3$ leaves.

option 1 may be warranted in some settings, but such a debate is beyond our scope. In this paper, we focus on option 2. Our main result is to prove that there exists a class of BD models that are identifiable based on lineage-through-time (LTT) data from an extant timetree. By identifiable, we mean that, within the space of rate functions we consider, each distinct BD model corresponds to one and only one likelihood function, and conversely. In fact, this space consists simply of piecewise constant rate functions, which are already widely used to fit BD models in practice.

Our results show that this class is identifiable once there are enough leaves in the extant tree, and we derive explicit lower bounds on the requisite number of samples. These bounds depend on a measure of parsimony of the underlying model class: They require that identifiable classes of birth–death rate functions do not oscillate unnaturally, in a sense that is made precise below. The same phenomenon has previously been observed in population genetics (17, 18), and our proofs are based in part on these earlier works.

1. Preliminaries

In this section, we define the BD model and introduce some key definitions.

Throughout the paper, n is used to denote the number of internal branching events, so that $n + 1$ is the number of leaves. We assume $n \geq 1$ and suppress explicit dependence on it when there is no risk of confusion. Given $n + 1$ sampled taxa, an extant timetree is a bifurcating tree that traces out the ancestry of the sample. Therefore, the extant timetree has n internal nodes that denote the times at which various taxa diverged from common ancestors. These are denoted $0 \leq \tau_n < \cdots < \tau_1$, where time runs backward from the present. As in LP (16), we assume that all $n + 1$ samples are collected at time $t = 0$. There is also a root node referred to as the origin that occurs at height $\tau_o < \infty$, when the process is assumed to have started. The height of the origin node is not resolvable from character data evolving along the tree since it is ancestral to the entire sample, so its value is conditioned on using prior information. An example of an extant timetree with three leaves is shown in Fig. 1.

Extant timetrees are assumed to be stochastically generated by a BD process (4, 14). This process has three parameters: two positive rate functions $\lambda : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ and $\mu : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ and an initial sampling fraction $\rho \in (0, 1]$. Here λ and μ represent the instantaneous rate per capita at which lineages are born and die going forward in time, and each lineage surviving to the present is sampled independently with probability ρ. Henceforth, we refer to different BD models by their corresponding parameter triples (λ, μ, ρ). Under the BD model with parameters (λ, μ, ρ), the density of an extant timetree is denoted $L^{(\lambda,\mu,\rho)}(\tau_1, \ldots, \tau_n)$. The precise form of $L^{(\lambda,\mu,\rho)}$ is not important for what follows, but can be found in Morlon et al. (ref. 5, equation 1). Note that the topology of the timetree is uninformative in this model; the likelihood depends only on the merger times τ_i.

Turning now to the concept of identifiability, let Θ be an arbitrary parameter space, and let L_θ denote a likelihood function parameterized by $\theta \in \Theta$. The statistical model $\mathcal{L}_\Theta = \{L_\theta : \theta \in \Theta\}$ is the image of Θ under L_θ, that is, the set of all possible likelihood functions that can be obtained from the parameter space Θ. If Θ is a set of BD parameters, we use the notation

$$\mathcal{B}_\Theta = \left\{ L^{(\lambda,\mu,\rho)} : (\lambda, \mu, \rho) \in \Theta \right\} \qquad [1]$$

to emphasize that we are focusing specifically on the BD model.

Definition 1 (identifiability): The statistical model $\mathcal{L}_\Theta = \{L_\theta : \theta \in \Theta\}$ is identifiable if $\theta \mapsto L_\theta$ is injective; that is, for all θ_1, $\theta_2 \in \Theta$, we have $L_{\theta_1} = L_{\theta_2} \implies \theta_1 = \theta_2$.

In the context of the BD model Eq. **1**, the statement "$(\lambda_1, \mu_1, \rho_1) = (\lambda_2, \mu_2, \rho_2)$" is understood to mean that $\rho_1 = \rho_2$ and that the corresponding rate functions are equal almost everywhere. Similarly, two density functions $L_{\theta_1}, L_{\theta_2}$ are considered equal if they differ on at most a set of zero Lebesgue measure.

If different parameters yield the same likelihood function, they cannot be distinguished using any amount of observable data. Identifiability is therefore the most minimal regularity condition one can place on a statistical model.

2. Results

In this section, we summarize LP's (16) results, prove that piecewise constant BD models are identifiable, and explore some additional corollaries and conjectures.

A. The Result of Louca and Pennell (16). The key quantity that underlies LP's (16) result is the so-called pulled (birth) rate function λ_p, which is defined to be the relative slope of the (deterministic) number of lineages through time. They show that the relative slope is equivalently expressed as

$$\lambda_p = \lambda \cdot (1 - E), \qquad [2]$$

where $E(t)$ is the probability that a lineage alive at time t has no descendants sampled at time 0. Then Eq. **2**. shows that the actual birth rate λ is "pulled" downward to obtain the function λ_p. The antiderivative of λ_p is denoted

$$\Lambda_p(\tau) = \int_0^\tau \lambda_p(u)\,\mathrm{d}u.$$

The function E satisfies the ordinary differential equation

$$\frac{\mathrm{d}E}{\mathrm{d}\tau} = \mu - (\lambda + \mu) \cdot E + \lambda E^2, \qquad [3]$$

with initial condition $E(0) = 1 - \rho$. The solution to Eq. **3**. is (5)

$$E(\tau) = 1 - \frac{e^{\int_0^\tau \lambda(u) - \mu(u)\,\mathrm{d}u}}{\rho^{-1} + \int_0^\tau \lambda(u) e^{\int_0^u \lambda(v) - \mu(v)\,\mathrm{d}v}\,\mathrm{d}u}. \qquad [4]$$

Note that E is continuous, even if λ and μ are not.

The pulled rate function completely characterizes the likelihood of an extant timetree. Specifically, by equation 34 of LP (16),

$$L^{(\lambda_p)}(\tau_1,\dots,\tau_n) \propto e^{-\Lambda_p(\tau_o)}\prod_{j=1}^{n}\lambda_p(\tau_j)e^{-\Lambda_p(\tau_j)}, \quad 0\le\tau_n\le\cdots\le\tau_1<\tau_o. \qquad [5]$$

By implication, any two BD parameter triples (λ_1,μ_1,ρ_1) and (λ_2,μ_2,ρ_2) that generate the same λ_p via Eq. **2** are indistinguishable. LP's (16) contribution is to show that this phenomenon emerges in a surprisingly general class of models. Restated in our notation, their main result is as follows:

Theorem (LP). *Given an extant timetree on $n+1$ taxa with origin τ_o, let $\mathcal{C}^1_+[0,\tau_o]$ denote the space of all functions that are strictly positive and continuously differentiable on $[0,\tau_o]$, and let*

$$\mathcal{U}=\left\{(\lambda,\mu,\rho):\lambda,\mu\in\mathcal{C}^1_+,\rho\in(0,1]\right\}$$

be the set of all BD parameterizations derived from this space. Then the BD model $\mathcal{B}_{\mathcal{U}}$ is unidentifiable.

Importantly, *Theorem (LP)* holds for any number of mergers n and also if ρ is fixed. LP's (16) proof is constructive and provides, for any given BD model, a set of infinitely many "congruent" models that all have the same likelihood. As LP (16) argue in their discussion, this result has disturbing implications for the reliability of statistical estimates obtained from BD models, which have been widely reported in phylogenetics, phylodynamics, paleogenetics, and related fields.

B. Piecewise Constant Models Are Identifiable. In this section, we state our main results.

Definition 2: Let

$$C^{\oplus K}_+[0,\tau_o]=\left\{\sum_{k=1}^{K}a_k\mathbf{1}_{[t_{k-1},t_k)}(t):\mathbf{a}\in\mathbb{R}^K_{>0},0=t_0<t_1<\cdots<t_K=\tau_o\right\}$$

be the set of all positive piecewise constant functions with K pieces defined on $[0,\tau_o]$.

Note from the definition that $C^{\oplus K}_+[0,\tau_o]$ encompasses all possible piecewise constant functions with K breakpoints. The location of the breakpoints can vary between models; we do not assume that all models are defined on a set of common breakpoints.

Next, we define the class of BD parameterizations that forms the basis of our identifiability proof. In the definition and in what follows, we assume that the sampling fraction $\rho\in(0,1]$ is a fixed, known parameter. This is necessary because if ρ is allowed to vary, then as noted in the Introduction, Stadler (14) has shown that even the constant-rates BD model is unidentifiable.

Definition 3: Let

$$\mathcal{I}_{K,\rho}=\left\{(\lambda,\mu,\rho):\lambda,\mu\in C^{\oplus K}_+[0,\tau_o]\right\}$$

be the space of all piecewise-constant BD parameterizations with rate functions in $C^{\oplus K}_+[0,\tau_o]$ and fixed sampling fraction $\rho\in(0,1]$.

The following is our main result:

Theorem 4. *If $n>8K$, then the BD model $\mathcal{B}_{\mathcal{I}_{K,\rho}}$ is identifiable.*

The proof of *Theorem 4* is rather technical and is provided in *Appendix A. Proof of Theorem 4*. For the reader's convenience, we outline the major steps here:

Sketch of proof. First, we establish (*Proposition 7*) the existence of a numerical "signature" that is associated with the likelihood function of an extant timetree in a phylogenetic BD model. Any two likelihoods that are equal possess the same signature; conversely, if two models have a different signature, then their likelihoods are different, and hence they are distinguishable from one another given infinite data. Moreover, this signature is determined entirely by the pulled rate function. Next, we show (*Proposition 8*) that if there are two pulled rate functions that have the same signature, then either 1) the pulled rate functions are equal or 2) the pulled rate functions must oscillate in a certain way. Finally, we prove (*Proposition 10* onward) that, under the condition $n>8K$ stated above, the pulled rate function of a piecewise constant BD model is incapable of oscillating in this way and moreover that distinct piecewise constant BD models have different pulled rate functions. Thus, any two distinct piecewise-constant BD models have different signatures. This implies that they have different likelihood functions—a fact that does not hold for the more general model class considered by LP (16). ■

Unpacking the result, it asserts that both the positions of the breakpoints (the vector $\mathbf{t}$ in *Definition 2*) and the levels of each piece (the vector $\mathbf{a}$ in the definition) of both $\lambda(t)$ and $\mu(t)$ are estimable given sufficient data. These breakpoints are not assumed to be shared between the two rate functions or indeed between any two functions in the piecewise constant model space considered by *Theorem 4*. If, as is common in practice, we do assume that $\lambda(t)$ and $\mu(t)$ are defined on the same set of breakpoints (while still allowing this set to vary between different parameterizations in $\mathcal{I}_{K,\rho}$), then easy modifications to the proof show that $n>4K$ suffices for identifiability.

Several extensions and conjectures follow naturally from *Theorem 4*. Since it is possible to uniformly approximate a regular function class over a compact set using step functions, identifiable BD models are in some sense dense in the space of all BD models. A prototypical result is as follows:

Theorem 5. *Let $\rho\in(0,1]$ be fixed, let*

$$\mathcal{F}=\left\{f\in\mathcal{C}^1_+[0,\tau_o]:\|f'\|_\infty<B\right\}$$

be the set of positive, continuously differentiable functions with bounded first derivative over $[0,\tau_o]$, and let $\Theta_{\mathcal{F},\rho}=\{(\lambda,\mu,\rho):\lambda,\mu\in\mathcal{F}\}$ denote the resulting BD parameter space. Then

1) *$\mathcal{B}_{\Theta_{\mathcal{F},\rho}}$ is unidentifiable; and*
2) *There exists a set of functions $\mathcal{G}$ defined over $[0,\tau_o]$ such that for any $\epsilon>0$,*
   a) $\sup_{f\in\mathcal{F}}\inf_{g\in\mathcal{G}}\|f-g\|_\infty<\epsilon$, and
   b) $\mathcal{B}_{\Theta_{\mathcal{G},\rho}}$ is identifiable if $n>8B\tau_o/\epsilon$.

Proof: The first claim follows from LP (16), because their congruence classes include smooth perturbations of constant-rate BD models. For the second one, if $f\in\mathcal{F}$, then

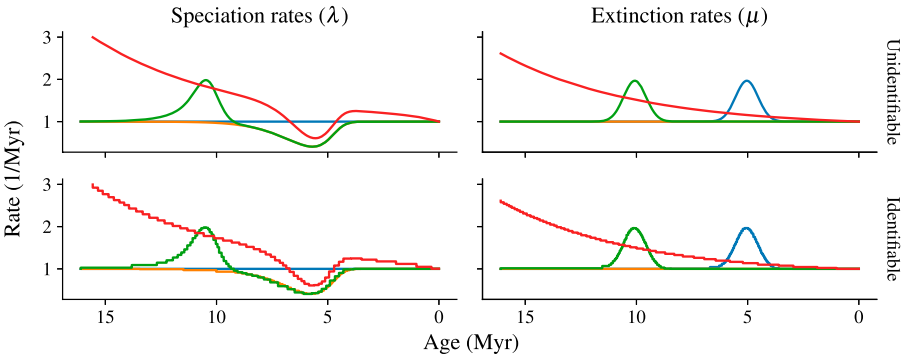

Fig. 2. Unidentifiable vs. identifiable BD models. *Top* row contains four indistinguishable models exhibited in LP's (16) figure 1. In *Bottom* row, we approximated these models using piecewise constant functions using $K = 50$ pieces. The models in *Bottom* row are identifiable given sufficiently many samples. All models are assumed to have the same ρ.

$$\begin{aligned}
&\left|\sum_{k=0}^{K-1} f(\tau_o k/K)\mathbf{1}_{[k/K,(k+1)/K)}(x/\tau_o) - f(x)\right| \\
&\quad\le \max_k \sup_{x/\tau_o\in[k/K,(k+1)/K)} |f(x) - f(\tau_o k/K)| \\
&\quad\le \max_k \sup_{x/\tau_o\in[k/K,(k+1)/K)} B|x - \tau_o k/K| \\
&\quad= B\tau_o/K.
\end{aligned}$$

Letting $K = B\tau_o/\epsilon$ yields the claim. ■

An obvious caveat to *Theorem 5* is that the sample size needed to have (provable) identifiability grows rapidly as $\epsilon \to 0$.

Another possible extension relates to estimating birth–death models using polynomials. Since constant functions are polynomials of degree zero, it is natural to conjecture that identifiability holds for higher degrees as well.

Conjecture 6. *Let $P_{d,+}^{(\oplus K)}[0,\tau_o]$ be the set of nonnegative, piecewise polynomials of order d with $K-1$ internal knots defined over $[0,\tau_o]$, and let $\Theta_{P_{d,+}^{(\oplus K)},\rho}[0,\tau_o]$ be the corresponding BD parameter space, again for fixed ρ. Then the BD model $\mathcal{B}_{\Theta_{P_{d,+}^{(\oplus K)},\rho}[0,\tau_o]}$ is identifiable if $n > 8K(1+d)$.*

Conjecture 6 would seem to imply that n grows with d, but this would be offset by having to use fewer pieces to obtain a good approximation. We are unable to prove *Conjecture 6* because substantial difficulties arise when trying to extend our proof technique to nonconstant functions. Specifically, we do not know how to bound the sign change complexity of spline-based BD models (see *Lemma 13*) except when $d = 0$.

3. Discussion

In this paper, we proved that piecewise-constant BD models are identifiable from extant timetrees with a sufficient number of tips. We also showed that, under mild assumptions, unidentifiable BD models of the type considered by LP (16) can be approximated to within arbitrary accuracy by identifiable BD models. Based on these results, we conjecture, but are unable to prove, that (piecewise) polynomial BD models are similarly identifiable.

In the short time since their publication, LP's (16) findings have generated considerable discussion (e.g., refs. 19–22), with some authors concluding that they "will be dispiriting to evolutionary scientists" seeking to understand the factors affecting speciation and extinction (20). Our results may serve to lift those spirits, while also illustrating potential subtleties that can arise when reasoning about a limiting concept like identifiability. For example, consider the BD models shown in Fig. 2. Fig. 2, *Top* row is reproduced from figure 1 of LP (16) and shows four color-coded BD models that all have the same pulled birth rate and hence the same likelihood function. In Fig. 2, *Bottom* row, we approximated these functions over the domain $[0, 16]$ using piecewise constant functions. By *Theorem 4*, these models can, in principle, be distinguished given a sufficiently large timetree. Is the underlying natural process that is modeled in Fig. 2 inferable from data? The answer seemingly depends on whether the researcher believes that the piecewise functions shown in Fig. 2, *Bottom* row can faithfully represent this process. If the researcher believes piecewise functions do faithfully represent the process, then the answer is yes. If the researcher believes continuous functions are better, then our methods so far extend only to the conclusions of *Theorem 5*. Empirically, we note that it would be nearly impossible to differentiate (using, say, a simple hypothesis test) between one of the $\mathcal{C}_+^1$ models in Fig. 2 and its corresponding $C_+^{\oplus 50}$ approximation on the basis of a realistically achievable amount of data.

An important point concerning our main result (*Theorem 4*) is that it establishes only a sufficient condition for identifiability. It does not imply that piecewise models are unidentifiable if n is below the stated bounds; in other words, we do not know whether this bound is sharp. In our view, the main message is that piecewise constant models are identifiable if at least $\mathcal{O}(K)$ tips are sampled. A related point concerns cases where the true model is piecewise constant with a small number of pieces, say K_0, but the modeler, who does not know the "true" K_0, fits a much larger model containing $K \gg K_0$ pieces. Our theory shows that the true model is identifiable in two senses: First, it can be distinguished from all other piecewise constant models containing at most K_0 pieces, using at least $8K_0$ samples; and second, it can be distinguished from among all models containing at most K pieces, using $n > 8K$ samples. From an estimation standpoint, there are clear advantages to the model containing only K_0 parameters, since the resulting estimates would have lower error. However, if the modeler is unaware of K_0 and chooses the number of pieces to be such that $K \approx n/8$, those estimates will necessarily be noisier, even if the model is technically identifiable. Finally, a related question of practical importance is a necessary condition for identifiability, as some applications might not have enough tips to have provable identifiability.

As this example indicates, practitioners should be careful not to overinterpret affirmative identifiability results as conclusive evidence that high-quality estimates can be obtained on real

problems. Even in identifiable models, it is often the case that significant regularization and/or prior information have to be incorporated to obtain sensible results (7, 19, 23–25). Having established identifiability, the next step would be to understand the finite-sample accuracy and rate of convergence of piecewise constant estimators in BD models. This is a challenging theoretical problem that will require new ideas and techniques. Fortunately, since several popular software packages (e.g., ref. 26) already implement the piecewise constant BD model, there are already many simulation studies in the literature to help guide the way. We recommend that researchers utilize simulations to understand the possibilities and limitations for fitting phylogenetic BD models to a specific dataset.

The reader may wonder whether our result is somehow a byproduct of the fact that we consider piecewise constant—hence discontinuous—rate functions, whereas in LP (16) they are assumed to be continuously differentiable. In our opinion, this is not the main driver. Indeed, we believe that (cf. *Conjecture 6*) identifiable parameter spaces consisting of smooth functions also exist. Provisionally, we suspect that these spaces are identifiable because they are finite dimensional and have fundamentally lower complexity (in the sense of *Definition 9*) compared to the nonparametric function space considered by LP (16). Were the conjecture true, it would not contradict LP's (16) result, because the construction they use to generate their congruence classes (specifically, the operator $S[S_o, f]$ defined by supplementary equation 75 in ref. 1) is not closed over simple function spaces like fixed-degree polynomials. In other words, even f is a spline, and $S[S_o, f]$ is not. Thus, while there are infinitely large congruence classes of alternative BD parameterizations that are indistinguishable, the conjecture asserts that the intersection between these classes and a sufficiently simple function space consists of at most a single element. LP (16) provide a heuristic argument supporting this conjecture in section S.3 of their supplement.

In follow-up work, Louca et al. (27) study a more general model where sampling is allowed to occur over time and show that similar unidentifiability results hold in that setting as well. The coalescent-based methods we used in this paper, which condition on a number of lineages sampled at the present, do not readily extend to this setting, so our results leave open the question of whether piecewise-constant identifiability holds in random sampling models as well. In section S.2.2 of their supplement, Louca et al. (27) assert that restricting to piecewise constant model spaces cannot possibly resolve identifiability issues; however, their argument is nonrigorous and based only on simulation evidence. Our results establish that piecewise constant models are in fact identifiable. Nevertheless, identifiability is fundamentally a mathematical property that may have little bearing on one's (in)ability to successfully carry out inference in real-world problems. More research is needed to better understand the circumstances under which this is in fact possible.

Appendix A. Proof of Theorem 4

Our proof derives from a general technique developed by Bhaskar and Song (18) for establishing identifiability of rate functions in coalescent-type models. We follow their method closely, reproducing their results where necessary for completeness of exposition.

To build the necessary connections between the BD and coalescent models, we first note that $L^{(\lambda_p)}$ in Eq. **5** can be rewritten as

$$L^{(\lambda_p)}(\tau_1, \dots, \tau_n) \propto \prod_{j=1}^{n} j\lambda_p(\tau_j) e^{-j[\Lambda_p(\tau_j) - \Lambda_p(\tau_{j+1})]}, \quad [6]$$

where we defined $\tau_{n+1} \equiv 0$. This is the likelihood of a coalescent-type pure death process, where the "effective population size" is $1/\lambda_p(\tau)$, and where the rate of dying (backward in time) when there are j remaining lineages in the tree is $(j-1)\lambda_p(\tau)$ instead of the usual $\binom{j}{2}\lambda_p(\tau)$.

Our strategy for establishing identifiability is to construct a vector of invariants which, for a sufficiently large sample size, uniquely identifies the pulled rate function λ_p. To that end, given any pulled rate function λ_p and sample size n, we form an associated moment vector $\mathbf{c}^{(\lambda_p)} \in \mathbb{R}^n$, with entries

$$c_j^{(\lambda_p)} = \int_0^{\tau_o} e^{-j\Lambda_p(\tau)}\, \mathrm{d}\tau, \quad 1 \le j \le n. \quad [7]$$

Proposition 7. *Suppose that $L^{(\lambda_p^{(1)})}$ and $L^{(\lambda_p^{(2)})}$ are equal almost everywhere. Then $\mathbf{c}^{(\lambda_p^{(1)})} = \mathbf{c}^{(\lambda_p^{(2)})}$.*

Proof: In this proof we refer to the likelihood function for multiple sample sizes, so we let $L_n^{(\lambda_p)}(\tau_{n1}, \dots, \tau_{nn})$ be the likelihood of an extant timetree with $n+1$ sampled tips, where the merger times are $\tau_o \ge \tau_{n1} \ge \tau_{n2} \ge \cdots \ge \tau_{nn} \ge 0$. Expectation of a functional $f : \mathbb{R}^n \to \mathbb{R}$ with respect to $L_n^{(\lambda_p)}$ is denoted by $\mathbb{E}_{\lambda_p} f$; by definition, if $L_n^{(\lambda_p^{(1)})} = L_n^{(\lambda_p^{(2)})}$ almost everywhere, then $\mathbb{E}_{\lambda_p^{(1)}} f = \mathbb{E}_{\lambda_p^{(2)}} f$ for all measurable f.

We use some results from Kamm et al. (28) on moments of the truncated coalescent process, replacing each occurrence of the coalescent rate $\binom{j}{2}/N_e(\tau)$ with its corresponding rate in the BD model, $(j-1)\lambda_p(\tau)$. The expected value $\mathbb{E}_{\lambda_p}(\tau_o - \tau_{n1}) = \tau_o - \mathbb{E}_{\lambda_p}\tau_{n1}$ is written in their notation as $f_{n+1}^{\tau_o}(n+1 \mid \mathcal{A}_{\tau_o}^{(\lambda_p)} = 1)$, where $\mathcal{A}_\tau^{(\lambda_p)}$ is the birth–death process analog of the coalescent ancestral process, i.e., a pure death process on $\{n+1, \dots, 1\}$, which begins at state $n+1$ and transitions from state $j+1$ to state j at rate $j\lambda_p(\tau)$. By formulas 3 and 5 of Kamm et al. (28), we have

$$\begin{aligned}
\tau_o - \mathbb{E}_{\lambda_p}\tau_{n1} &= f_{n+1}^{\tau_o}(n+1 \mid \mathcal{A}_{\tau_o}^{(\lambda_p)} = 1) \\
&= \frac{f_{n+1}^{\tau_o}(n+1)}{\mathbb{P}_{n+1}(\mathcal{A}_{\tau_o}^{(\lambda_p)} = 1)} \\
&= \frac{\tau_o - \sum_{k=1}^{n} \frac{k}{n+1} f_{n+1}^{\tau_o}(k)}{\mathbb{P}_{n+1}(\mathcal{A}_{\tau_o}^{(\lambda_p)} = 1)},
\end{aligned}$$

where $f_{n+1}^{\tau_o}(k)$ is defined below. The quantity $\mathbb{P}_{n+1}(\mathcal{A}_{\tau_o}^{(\lambda_p)} = 1)$ is the probability that the unconditioned birth–death process reaches a common ancestor before time τ_o, meaning it is exactly the normalizing constant in Eq. **6**. Rearranging the preceding display and defining $d_n^{(\lambda_p)} = \mathbb{P}_{n+1}(\mathcal{A}_{\tau_o}^{(\lambda_p)} = 1)$, we obtain

$$\sum_{k=1}^{n} \frac{k}{n+1} f_{n+1}^{\tau_o}(k) = \tau_o[1 - d_n^{(\lambda_p)}] + d_n^{(\lambda_p)} \mathbb{E}_{\lambda_p}\tau_{n1}. \quad [8]$$

By lemma 3.3 of Kamm et al. (28), the summands in Eq. **8** are given by

$$f_{n+1}^{\tau_o}(k) = \sum_{m=2}^{n+1} W_{km}^{(n+1)} c_{m-1}^{(\lambda_p)}, \quad [9]$$

where the vector $\mathbf{c}^{(\lambda_p)}$ was defined in Eq. **7**, and the matrix $\mathbf{W}^{(n)}$ was derived by Polanski and Kimmel (29) in the case of Kingman's coalescent. In *Appendix B. Computation of the Matrix $\boldsymbol{W}^{(n)}$ for*

the BD Model, we derive a modified form of this matrix that is appropriate for use with the BD model.

Now from Eq. **5**, we have

$$L_m^{(\lambda_p)}(\tau_1,\ldots,\tau_m) \propto L_n^{(\lambda_p)}(\tau_o,\ldots,\tau_o,\tau_1,\ldots,\tau_m)$$

for any $1 \le m \le n$. Thus, given any $L_n^{(\lambda_p)}$, we may use the above procedure to calculate the moment vector

$$\mathbf{e}^{(\lambda_p)} = (\mathbb{E}_{\lambda_p}\tau_{11}, \mathbb{E}_{\lambda_p}\tau_{21}, \ldots, \mathbb{E}_{\lambda_p}\tau_{n1})^{\mathsf{T}}.$$

Define the lower-triangular matrix $\mathbf{B} = (b_{ij}) \in \mathbb{R}^{n\times n}$ to have entries

$$b_{ij} = \sum_{b=1}^{i} \frac{b}{i+1} W_{bj}^{(i+1)}, \quad 1 \le i \le n, \quad 2 \le j \le i+1,$$

where the second axis of $\mathbf{B}$ is indexed in the same manner as $\mathbf{W}^{(n)}$. In *Appendix B. Computation of the Matrix $\boldsymbol{W}^{(n)}$ for the BD Model*, we derive a closed-form expression for the entries of $\mathbf{B}$, which shows in particular that the diagonal entries $b_{i,i+1} = (-1)^{i+1}$. Therefore, $\mathbf{B}$ is invertible, so that by Eq. **8**,

$$\mathbf{c}^{(\lambda_p)} = \mathbf{B}^{-1}\left(\tau_o[\mathbf{I} - \mathrm{diag}(\mathbf{d}^{(\lambda_p)})] + \mathrm{diag}(\mathbf{d}^{(\lambda_p)})\mathbf{e}^{(\lambda_p)}\right). \quad \textbf{[10]}$$

Finally, suppose that $L_n^{(\lambda_p^{(1)})}$ and $L_n^{(\lambda_p^{(2)})}$ are two BD model likelihoods that are equal almost everywhere. Then there exists

$$0 = t_0 < t_1 < \cdots < t_K = \tau_o$$

such that $L_n^{(\lambda_p^{(1)})} - L_n^{(\lambda_p^{(2)})}$ is continuous on open rectangles of the form

$$\mathcal{R} = (t_{i_1}, t_{i_1+1}) \times \cdots \times (t_{i_n}, t_{i_n+1}) \subset \mathbb{R}^n$$

and equals zero almost everywhere on each such $\mathcal{R}$. Therefore, the preimage

$$(L_n^{(\lambda_p^{(1)})} - L_n^{(\lambda_p^{(2)})})^{-1}(\mathbb{R}^n \backslash \{\mathbf{0}\}) \cap \mathcal{R}$$

is an open set of zero measure; the only such set is $\emptyset$. Hence, $L_n^{(\lambda_p^{(1)})} = L_n^{(\lambda_p^{(2)})}$ everywhere on $\mathcal{R}$. In particular, this implies that for all $1 \le m \le n$, the BD likelihoods $L_m^{(\lambda_p{}^{(i)})}$ are equal almost everywhere on $\mathbb{R}^m$. Therefore, the vectors $\mathbf{d}^{(\lambda_p{}^{(i)})}$ and $\mathbf{e}^{(\lambda_p{}^{(i)})}$, which are defined entirely in terms integrals of $L_m^{(\lambda_p{}^{(i)})}$, $1 \le m \le n$, are equal for $i = 1, 2$. Eq. **10** then implies that $\mathbf{c}^{(\lambda_p^{(1)})} = \mathbf{c}^{(\lambda_p^{(2)})}$. ■

Contrapositively, if $\mathbf{c}^{(\lambda_p^{(1)})} \neq \mathbf{c}^{(\lambda_p^{(2)})}$, then $L^{(\lambda_p^{(1)})}$ and $L^{(\lambda_p^{(2)})}$ differ on a set of positive measure. The rest of the proof amounts to showing that if $\lambda_p^{(1)}$ and $\lambda_p^{(2)}$ are generated by piecewise constant BD models, and n is sufficiently large, then they have different moment vectors.

The next theorem is restated for completeness.

Theorem (Generalized Rule of Signs) (18, 30). *Let $f : \mathcal{D} \to \mathbb{R}$ be a piecewise-continuous function defined on some domain $\mathcal{D} \subset \mathbb{R}$, which is not identically zero and has a finite number $\sigma(f)$ of sign changes. Then the function*

$$G(x) = \int_{\mathcal{D}} f(t)e^{-tx}\,\mathrm{d}t$$

has at most $\sigma(f)$ zeros in $\mathbb{R}$ (counted with multiplicity).

Informally, f is said to have a sign change any time it crosses zero, including by jump discontinuities. For a precise statement, refer to definition 3 of Bhaskar and Song (18).

Given any pulled rate function λ_p, we define its time-rescaled rate function

$$\tilde{\lambda}_p(x) = \lambda_p({\Lambda_p}^{-1}(x)), \quad 0 \le x < \Lambda_p(\tau_o).$$

This transformation is invertible, since if

$$S_{\tilde{\lambda}_p}(t) = \int_0^t \left[\tilde{\lambda}_p(u)\right]^{-1}\mathrm{d}u = {\Lambda_p}^{-1}(t),$$

then $\tilde{\lambda}_p(S_{\tilde{\lambda}_p}^{-1}(t)) = \lambda_p(t)$. Hence,

$$\tilde{\lambda}_p^{(1)} = \tilde{\lambda}_p^{(2)} \iff \lambda_p^{(1)} = \lambda_p^{(2)}.$$

Then the entries of $\mathbf{c}^{(\lambda_p)}$ can be written as

$$c_j^{(\lambda_p)} = \int_0^{\Lambda_p(\tau_o)} \left[\tilde{\lambda}_p(x)\right]^{-1} e^{-jx}\,\mathrm{d}x.$$

Proposition 8. *Suppose that $\lambda_p^{(1)}$ and $\lambda_p^{(2)}$ are two pulled rate functions for which $\mathbf{c}^{(\lambda_p^{(1)})} = \mathbf{c}^{(\lambda_p^{(2)})} \in \mathbb{R}^n$. Then either $\lambda_p^{(1)} = \lambda_p^{(2)}$ or $[\tilde{\lambda}_p^{(1)}]^{-1} - [\tilde{\lambda}_p^{(2)}]^{-1}$ has at least $n-1$ sign changes over the shared domain of $[\tilde{\lambda}_p^{(1)}]^{-1}$ and $[\tilde{\lambda}_p^{(2)}]^{-1}$.*

Proof: Suppose that $\lambda_p^{(1)} \neq \lambda_p^{(2)}$. Then $[\tilde{\lambda}_p^{(1)}]^{-1} - [\tilde{\lambda}_p^{(2)}]^{-1}$ is not identically zero. Assume without loss of generality that $\Lambda_p^{(1)}(\tau_o) \le \Lambda_p^{(2)}(\tau_o)$ so the shared domain of $[\tilde{\lambda}_p^{(1)}]^{-1}$ and $[\tilde{\lambda}_p^{(2)}]^{-1}$ is $[0, \Lambda_p^{(1)}(\tau_o))$. Consider the integral transform given by

$$G(z) = \int_{x=0}^{\Lambda_p^{(2)}(\tau_o)} \left[u_1(x) - \frac{1}{\tilde{\lambda}_p^{(2)}(x)}\right] e^{-zx}\,dx,$$

where

$$u_1(x) = \begin{cases} [\tilde{\lambda}_p^{(1)}(x)]^{-1} & x \in [0, \Lambda_p^{(1)}(\tau_o)) \\ 0 & \text{otherwise.} \end{cases}$$

The supposition implies that $G(z)$ has zeros at $z = 1, \ldots, n$. By the generalized rule of signs, this implies that $u_1 - [\tilde{\lambda}_p^{(2)}]^{-1}$ has at least n sign changes on $[0, \Lambda_p^{(2)}(\tau_o))$. There is at most one sign change on the interval $[\Lambda_p^{(1)}(\tau_o), \Lambda_p^{(2)}(\tau_o))$, caused by a possible jump at $\Lambda_p^{(1)}(\tau_o)$. This implies that $u_1 - [\tilde{\lambda}_p^{(2)}]^{-1}$ has at least $n-1$ sign changes on $[0, \Lambda_p^{(1)}(\tau_o))$. ■

Based on the preceding result, we define the following complexity measure on BD model spaces. This is an adaptation of definition 4 in Bhaskar and Song (18) to our setting. In the definition, the notation ${\lambda_p}^{(\theta)}$ is used to denote the pulled rate function corresponding to a particular BD parameterization $\theta = (\lambda, \mu, \rho)$.

Definition 9 (pulled sign change complexity): Let Θ be a set of BD models, and let $\mathcal{G}$ be the set of all functions defined by the condition

$$g \in \mathcal{G} \iff \exists \theta_1, \theta_2 \in \Theta, a \ge 0 \text{such that}$$
$$g(x) = \left[\tilde{\lambda}_p{}^{(\theta_1)}(x)\right]^{-1} - \left[\tilde{\lambda}_p{}^{(\theta_2)}(x-a)\right]^{-1},$$

where the domain of each such function is

$$\mathrm{dom}(g) = \left[\max\{0, a\}, \min\{{\Lambda_p}^{(1)}(\tau_o), a + {\Lambda_p}^{(2)}\}\right).$$

The pulled sign change complexity of Θ is defined as

$$\mathscr{S}_p = \sup\{\sigma(g) : g \in \mathcal{G}\},$$

where $\sigma(g)$ denotes the number of sign changes of g.

In the calculation of pulled sign change complexity, we find the number of sign changes for each candidate function g. Each g consists of the difference of two time-rescaled and inverted pulled rate functions; one of the two pulled rate functions can be shifted by a units in the positive direction. Having a large number of sign changes indicates that at least one of the two models has many increasing and decreasing periods. Bounding the complexity of the model class Θ is tantamount to requiring that the birth and death rates do not oscillate in such a way. This is a sort of parsimony assumption since, in the extreme, the functions cannot oscillate at all and must be constant.

Using *Proposition 8* and the preceding definition, we immediately have the following sample size criterion for the identifiability of BD models:

Proposition 10. *Suppose that $\mathscr{S}_p(\Theta) \leq S$ and that the mapping $\theta \mapsto {\lambda_p}^{(\theta)}$ is injective over Θ. Then $\mathcal{B}_\Theta$ is identifiable if $n > S+1$.*

Proposition 10 is a general result that holds for any BD model class Θ. However, Θ must be chosen so that $\theta \mapsto \lambda_p^{(\theta)}$ is injective and $\mathscr{S}_p(\Theta) \leq S$ for a given S. To prove *Theorem 4*, it remains to establish these properties when $\Theta = \mathcal{I}_{K,\rho}$ and $S = 8K - 1$. Injectivity is shown in *Proposition 11*, and the sign change complexity is bounded in *Lemmas 12* and *13*.

Recall that $\tilde{\lambda}_p(x) = \lambda_p({\Lambda_p}^{-1}(x))$ for $x \in [0, \Lambda_p(\tau_o))$. By supplemental equation 9 of LP (16),

$$\frac{\mathrm{d}\lambda_p}{\mathrm{d}t} = \lambda_p\left(\frac{1}{\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}t} - \mu + \lambda E\right).$$

Hence,

$$\begin{aligned}\frac{\mathrm{d}\tilde{\lambda}_p}{\mathrm{d}x} &= \frac{\mathrm{d}\lambda_p}{\mathrm{d}t}(\Lambda_p^{-1}(x)) \times \frac{\mathrm{d}\Lambda_p^{-1}}{\mathrm{d}x}(x) \\ &= \frac{1}{\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}t} - \mu + \lambda E\Big|_{t=\Lambda_p^{-1}(x)},\end{aligned} \quad \textbf{[11]}$$

where in the second equality we used

$$\frac{\mathrm{d}\Lambda_p^{-1}}{\mathrm{d}x} = \frac{1}{\lambda_p(\Lambda_p^{-1}(x))} = [\tilde{\lambda}_p(x)]^{-1}.$$

Now by Eq. **2**,

$$\lambda E|_{t=\Lambda_p^{-1}(x)} = -\tilde{\lambda}_p(x) + \lambda(\Lambda_p^{-1}(x)). \quad \textbf{[12]}$$

If λ and μ are constant, then $\mathrm{d}\lambda/\mathrm{d}t = 0$, and we obtain from Eqs. **11** and **12** the first-order ordinary differential equation

$$\begin{aligned}\frac{\mathrm{d}\tilde{\lambda}_p}{\mathrm{d}x} &= \lambda - \mu - \tilde{\lambda}_p \\ \tilde{\lambda}_p(0) &= \rho\lambda.\end{aligned} \quad \textbf{[13]}$$

The solution to this differential equation is

$$\tilde{\lambda}_p(x) = (\lambda - \mu)(1 - e^{-x}) + \rho\lambda e^{-x}, \quad x \in [0, \Lambda_p(\tau_o)). \quad \textbf{[14]}$$

More generally, if λ and μ are constant over some interval $[t, t')$, then

$$\begin{aligned}\tilde{\lambda}_p(x) = (\lambda - \mu)(1 - e^{-(x-\Lambda_p(t))}) + \lambda_p(t)e^{-(x-\Lambda_p(t))}, \\ x \in [\Lambda_p(t), \Lambda_p(t')).\end{aligned} \quad \textbf{[15]}$$

Proposition 11. *Let θ_1, θ_2 be two different models in $\mathcal{I}_{K,\rho}$ with pulled rate functions $\lambda_p^{(1)}$ and $\lambda_p^{(2)}$. Then $\lambda_p^{(1)} \neq \lambda_p^{(2)}$.*

Proof: Let $(\lambda_1, \mu_1) \neq (\lambda_2, \mu_2)$ be two different models in $\mathcal{I}_{K,\rho}$. Then there is a nonempty interval $[t, t') \subset [0, \tau_o]$ such that

1) $(\lambda_1(s), \mu_1(s)) = (\lambda_2(s), \mu_2(s))$ for all $0 < s < t$; and
2) $\lambda_1, \mu_1, \lambda_2, \mu_2$ are all constant over $[t, t')$ and $(\lambda_1(s), \mu_1(s)) \neq (\lambda_2(s), \mu_2(s))$ for all $t \leq s < t'$.

(Note that we could have $t = 0$, in which case condition 1 becomes vacuous.) To show $\lambda_p^{(1)} \neq \lambda_p^{(2)}$, it is sufficient to show that $\tilde{\lambda}_p^{(1)} \neq \tilde{\lambda}_p^{(2)}$. We assume that $\Lambda_p^{(1)}(t') = \Lambda_p^{(2)}(t')$, since if not the conclusion is immediate. By Eq. **15**, for all $x \in [\Lambda_p^{(1)}(t), \Lambda_p^{(1)}(t'))$, we have

$${\tilde{\lambda}_p}^{(2)}(x) - {\tilde{\lambda}_p}^{(1)}(x) = c_1 e^{-[x-\Lambda_p^{(1)}(t)]} + c_2,$$

where

$$\begin{aligned}c_1 &= \lambda_p^{(1)}(t) - \lambda_p^{(2)}(t) - \lambda_2 + \mu_2 + \lambda_1 - \mu_1 \\ c_2 &= \lambda_2 - \mu_2 - \lambda_1 + \mu_1.\end{aligned}$$

Suppose $c_2 = 0$. Let $\varepsilon = E^{(1)}(t) = E^{(2)}(t) \in (0, 1)$, where the equality follows from condition 1 and the facts that a) $E(0) = \rho$ across all models, and b) $E(t)$ is continuous (cf. Eq. **4**). Then $c_1 = \lambda_p^{(1)}(t) - \lambda_p^{(2)}(t) = [\lambda_1(t) - \lambda_2(t)](1 - \varepsilon)$. If $c_1 = 0$, then this would contradict condition 2. ■

Remark: The preceding result makes crucial use of the fact that all models in $\mathcal{I}_{K,\rho}$ are constrained to have the same sampling fraction ρ. Without this assumption, *Proposition 11* would not even hold for $K = 1$ (14).

Next, we bound $\mathscr{S}_p(\mathcal{I}_{K,\rho})$. First, let

$$\mathcal{S}_K = \bigcup_{\rho \in (0,1]} \mathcal{I}_{K,\rho}$$

be the space of all K-piecewise constant BD models with unconstrained sampling proportions. As remarked above, this space is not identifiable, since in particular *Proposition 11* does not hold for it. Nevertheless, it follows directly from *Definition 9* that $\mathscr{S}_p(\mathcal{I}_{K,\rho}) \leq \mathscr{S}_p(\mathcal{S}_K)$, so bounding the pulled sign change complexity of $\mathcal{S}_K$ is all that is required for our purposes.

We first show that $\mathscr{S}_p(\mathcal{S}_K)$ can be bounded in terms of the simpler quantity $\mathscr{S}_p(\mathcal{S}_1)$.

Lemma 12. *The pulled sign change complexity of $\mathcal{S}_K$ is bounded by the pulled sign change complexity of $\mathcal{S}_1$ as*

$$\mathscr{S}_p(\mathcal{S}_K) \le (4K-1) + 4K\mathscr{S}_p(\mathcal{S}_1).$$

Proof: Let ${\lambda_p}^{(i)}$ be the pulled rate function corresponding to $(\lambda_i, \mu_i, \rho_i)$ for $i = 1, 2$. According to *Definition 9*, we need to bound all sign changes of

$$\left[\tilde{\lambda}_p{}^{(1)}(x)\right]^{-1} - \left[\tilde{\lambda}_p{}^{(2)}(x-a)\right]^{-1} \qquad \textbf{[16]}$$

over the domain $x \in [m, M)$, where

$$m = \max\{0, a\}$$
$$M = \min\{{\Lambda_p}^{(1)}(\tau_o), a + {\Lambda_p}^{(2)}(\tau_o)\}.$$

Enlarging the domain of Eq. **16** can only increase the number of sign changes, and the largest possible domain occurs when $a = 0$ and ${\Lambda_p}^{(1)}(\tau_o) = {\Lambda_p}^{(2)}(\tau_o)$, so we assume these conditions hold for the rest of the proof.

If $\lambda_i, \mu_i \in C_+^{\oplus K}$, then we can place them onto a common set of $2K$ breakpoints

$$0 = t_0^{(i)} < t_1^{(i)} < \cdots < t_{2K}^{(i)} = \tau_o.$$

Let

$$\mathcal{X} = \left\{ {\Lambda_p}^{(i)}(t_k^{(i)}) : 1 \le i \le 2, 0 \le k \le 2K \right\},$$

and sort the points in $\mathcal{X}$ to form a partition

$$0 = x_0 < \cdots < x_{4K} = {\Lambda_p}^{(1)}(\tau_o) = {\Lambda_p}^{(2)}(\tau_o).$$

Allowing for possible jump discontinuities at $x_1, x_2, \ldots, x_{4K-1}$, the number of sign changes of Eq. **16** is at most $4K-1$ plus the number of sign changes on each interval (x_j, x_{j+1}).

For each i and j, there exists an integer $0 \le k(i,j) < 2K$ such that

$$(x_j, x_{j+1}) \subset \left(\Lambda_p^{(i)}(t_{k(i,j)}^{(i)}), \Lambda_p^{(i)}(t_{k(i,j)+1}^{(i)})\right), \quad i = 1, 2.$$

Therefore, there exists a BD parameterization $\theta_{ij} = (\lambda_{ij}, \mu_{ij}, \rho_{ij}) \in \mathcal{S}_1$ such that

$$\lambda_p^{(\theta_{ij})}(s - t_{k(i,j)}^{(i)}) = \lambda_p^{(i)}(s), \quad s \in (t_{k(i,j)}^{(i)}, t_{k(i,j)+1}^{(i)});$$

concretely, the initial sampling fraction is

$$\rho_{ij} = 1 - E^{(i)}(t_{k(i,j)}^{(i)}).$$

Then

$$\begin{aligned}\left[\tilde{\lambda}_p{}^{(i)}(x)\right]^{-1} &= \left[\lambda_p(I^{(i)}(x))\right]^{-1} \\ &= \left[\tilde{\lambda}_p{}^{(\theta_{ij})}\left(x - \Lambda_p^{(i)}(t_{k(i,j)}^{(i)})\right)\right]^{-1}, \quad i = 1, 2.\end{aligned}$$

So within (x_j, x_{j+1}), the number of sign changes of $[\tilde{\lambda}_p^{(1)}(x)]^{-1} - [\tilde{\lambda}_p^{(2)}(x)]^{-1}$ is at most the number of sign changes of

$$\left[\tilde{\lambda}_p{}^{(\theta_{1j})}\left(x - \Lambda_p^{(1)}(t_{k(1,j)}^{(1)})\right)\right]^{-1} - \left[\tilde{\lambda}_p{}^{(\theta_{2j})}\left(x - \Lambda_p^{(2)}(t_{k(i,j)}^{(2)})\right)\right]^{-1},$$

which is bounded above by $\mathscr{S}_p(\mathcal{S}_1)$. Hence, the number of sign changes is at most $(4K-1) + 4K\mathscr{S}_p(\mathcal{S}_1)$. ■

We conclude the proof by bounding $\mathscr{S}(\mathcal{S}_1)$.

Lemma 13. *Let $(\lambda_1, \mu_1, \rho_1), (\lambda_2, \mu_2, \rho_2) \in \mathcal{S}_1$, with corresponding pulled rate functions $\lambda_p^{(1)}$ and $\lambda_p^{(2)}$, and let*

$$g(x) = \left[\tilde{\lambda}_p^{(1)}(x)\right]^{-1} - \left[\tilde{\lambda}_p^{(2)}(x-a)\right]^{-1},$$

where $a \ge 0$ is arbitrary and the domain of g is as indicated in Definition 9. Then $\sigma(g) \le 1$.

Proof: We have

$$g(x) = \frac{\tilde{\lambda}_p^{(2)}(x-a) - \tilde{\lambda}_p^{(1)}(x)}{\tilde{\lambda}_p^{(1)}(x)\tilde{\lambda}_p^{(2)}(x-a)},$$

so the number of sign changes of g is at most the number of zeros of $\tilde{\lambda}_p^{(2)}(x-a) - \tilde{\lambda}_p^{(1)}(x)$. By Eq. **14**, the function $\tilde{\lambda}_p^{(2)}(x-a) - \tilde{\lambda}_p^{(1)}(x)$ has the form $c_1 e^{-x} + c_2$ for some c_1, c_2 that depend on λ_i, μ_i, ρ_i, and a. Since this function is always monotone, $\tilde{\lambda}_p^{(2)}(x-a) - \tilde{\lambda}_p^{(1)}(x)$ and hence g has at most one zero. ■

By Lemma 13, $\mathscr{S}_p(\mathcal{S}_1) \le 1$, so that by *Lemma 12*,

$$\mathscr{S}_p(\mathcal{S}_K) \le 8K - 1.$$

Finally, *Theorem 4* follows from *Proposition 10*.

Appendix B. Computation of the Matrix $\mathbf{W}^{(n)}$ for the BD Model

In this section we derive the matrix $\mathbf{W}^{(n)} \in \mathbb{Q}^{(n-1)\times(n-1)}$ under the BD model. Starting from equation 1 of Polanski and Kimmel (29), which originally appeared in Griffiths and Tavaré (31), we have

$$\begin{aligned} q_{nb} &= \frac{\sum_{k=2}^{n} \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} k\mathbb{E}(S_k)}{\sum_{k=2}^{n} k\mathbb{E}(S_k)} \\ &= \frac{\frac{(n-b-1)!(b-1)!}{(n-1)!}\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1}\mathbb{E}(S_k)}{\sum_{k=2}^{n} k\mathbb{E}(S_k)} \end{aligned}, \qquad \textbf{[17]}$$

where S_k is the expected amount of time spent at level K in a coalescent tree. The S_k are defined as the differences $S_k = T_k - T_{k+1}$, where T_k is the height of the Kth coalescent event, and $T_{n+1} \equiv 0$. By equation 5 of Polanski et al. (32),

$$\mathbb{E}(T_k) = \sum_{j=k}^{n} A_{kj}^{(n)} c_j,$$

where c_j is the expected time to first coalescence in a sample of size j, defined in Eq. 7 of the main text for the phylogenetic BD model, and $\mathbf{A}^{(n)}$ is a matrix of combinatorial coefficients that have to be modified from their original definition (equation 6 of ref. 29) to reflect the coalescence rate of the BD process:

$$A_{kj}^{(n)} = \prod_{\ell=k, \ell\neq j}^{n} \frac{\ell - 1}{\ell - j}, \quad 2 \le k \le j \le n,$$

and zero otherwise. From the definition, we see that

$$A^{(n)}_{k+1,j} = \frac{k-j}{k-1} A^{(n)}_{kj}, \quad k \geq 2,$$

and therefore, following equation 51 of Polanski et al. (32), we have

$$\begin{aligned}
\mathbb{E}S_k &= \sum_{j=k}^{n} A^{(n)}_{kj} c_j - \sum_{j=k+1}^{n} A^{(n)}_{k+1,j} c_j \\
&= A^{(n)}_{kk} c_k + \sum_{j=k+1}^{n} \frac{j-1}{k-1} A^{(n)}_{kj} c_j. \\
&= \sum_{j=k}^{n} \frac{j-1}{k-1} A^{(n)}_{kj} c_j.
\end{aligned}$$

Inserting this expression into Eq. **17** and simplifying, we obtain

$$\begin{aligned}
q_{nb} &= \frac{\frac{(n-b-1)!(b-1)!}{(n-1)!} \sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1} \sum_{j=k}^{n} \frac{j-1}{k-1} A^{(n)}_{kj} c_j.}{\sum_{k=2}^{n} k\mathbb{E}(S_k)} \\
&= \frac{\sum_{j=2}^{n}(j-1)c_j \sum_{k=2}^{j} \frac{(n-b-1)!(b-1)!}{(n-1)!}\binom{n-k}{b-1} k A^{(n)}_{kj}}{\sum_{k=2}^{n} k\mathbb{E}(S_k)}.
\end{aligned}$$

The matrix $\mathbf{W}^{(n)}$ is defined to be

$$W^{(n)}_{bj} = (j-1) \sum_{k=2}^{j} \frac{(n-b-1)!(b-1)!}{(n-1)!} \binom{n-k}{b-1} k A^{(n)}_{kj}. \quad \textbf{[18]}$$

$\mathbf{c}$ is defined so that $q_{nb} \propto \mathbf{W}^{(n)}\mathbf{c}$.

A. Recursion for $\mathbf{W^{(n)}}$. Although we do not require it in this paper, following Polanski and Kimmel (29), we derived a recursion for computing the entries of $\mathbf{W}^{(n)}$. Using Zeilberger's algorithm (33), we obtain

$$\begin{aligned}
W^{(n)}_{b,j} = \Big[&(bj-(j-2)(n+1))\left[(b(j+3)+j(-2j+n+6)\right. \\
&\left.-4n-1\right] W^{(n)}_{b,j-1} - (b-j+4)(j-n-2)\, W^{(n)}_{b,j-2}\Big] \\
&\Big/ \left\{(j-2)\left[\left(j^2-7\right)(n+1) - b(j-1)(j+3)\right]\right\}
\end{aligned}$$

with base cases

$$\begin{aligned}
W^{(n)}_{b,2} &= 2 \\
W^{(n)}_{b,3} &= n - 3b + 1.
\end{aligned}$$

In contrast to the case of Kingman's coalescent, the denominator in the above recursion can be zero for certain settings of n, b, and j. (For example, $n = 15, j = 5, b = 9$.) In that case, an alternative, one-term recurrence is also available:

$$W^{(n)}_{b,j} = \frac{(j-b-3)[j(b-n-1)+2(n+1)]}{(j-2)[b(j-1)-(j-3)(n+1)]} W^{(n)}_{b,j-1}, j \geq 3.$$

(Observe that the denominators of the two recursions are not simultaneously zero unless $j = 2$.)

B. The Matrix B. From the binomial identity

$$\sum_{k=0}^{m} \binom{m}{k} \Big/ \binom{n}{k} = \frac{n+1}{n+1-m},$$

we obtain for $k \geq 2$,

$$\begin{aligned}
\sum_{b=1}^{n-1} \frac{\binom{n-k}{b-1}}{\binom{n-1}{b}} &= \sum_{b=1}^{n-1} \frac{\binom{n-k+1}{b} - \binom{n-k}{b}}{\binom{n-1}{b}} \\
&= \sum_{b=0}^{n-1} \frac{\binom{n-k+1}{b}}{\binom{n-1}{b}} - \sum_{b=0}^{n-1} \frac{\binom{n-k}{b}}{\binom{n-1}{b}} \\
&= \sum_{b=0}^{n-k+1} \frac{\binom{n-k+1}{b}}{\binom{n-1}{b}} - \sum_{b=0}^{n-k} \frac{\binom{n-k}{b}}{\binom{n-1}{b}} \\
&= \frac{n}{n-(n-k+1)} - \frac{n}{n-(n-k)} \\
&= \frac{n}{k(k-1)}.
\end{aligned}$$

Inserting this into Eq. **18** and simplifying, we obtain

$$\sum_{b=1}^{n-1} \frac{b}{n} W^{(n)}_{bj} = (j-1) \sum_{k=2}^{j} \frac{A^{(n)}_{kj}}{k-1}.$$

Furthermore,

$$\begin{aligned}
\sum_{k=2}^{j} \frac{A^{(n)}_{kj}}{k-1} &= \sum_{k=2}^{j} \frac{\prod_{\ell=k,\ell\neq j}^{n} \frac{\ell-1}{\ell-j}}{k-1} \\
&= \prod_{\ell=j+1}^{n} \frac{\ell-1}{\ell-j} \sum_{k=2}^{j} \frac{\prod_{\ell=k}^{j-1} \frac{\ell-1}{\ell-j}}{k-1} \\
&= \prod_{\ell=j+1}^{n} \frac{\ell-1}{\ell-j} \left(\frac{1}{j-1} + \sum_{k=2}^{j-1} \frac{\frac{(j-2)!}{(k-2)!} \times (-1)^{j+k}}{(k-1)(j-k)!} \right) \\
&= \prod_{\ell=j+1}^{n} \frac{\ell-1}{\ell-j} \left(\frac{1}{j-1} + \frac{1}{j-1} \sum_{k=2}^{j-1} \binom{j-1}{k-1} \right. \\
&\qquad \left. (-1)^{j+k} \right) \\
&= \prod_{\ell=j+1}^{n} \frac{\ell-1}{\ell-j} \times \frac{(-1)^j}{j-1} \\
&= \frac{(n-1)!}{(j-1)!(n-j)!} \times \frac{(-1)^j}{j-1} \\
&= \frac{(-1)^j}{j-1} \binom{n-1}{j-1}.
\end{aligned}$$

We conclude that

$$\sum_{b=1}^{n-1} \frac{b}{n} W^{(n)}_{bj} = (-1)^j \binom{n-1}{j-1}.$$

Data, Materials, and Software Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. We thank Andy Magee, Sebastian Prillo, Ed Ionides, and the two referees for comments that improved the manuscript. All errors are our own. This research was supported by NSF Grants DMS-2052653 and DMS-1646108.

1. W. Feller, Die grundlagen der volterraschen theorie des kampfes ums dasein in wahrscheinlichkeitstheoretischer behandlung. *Acta Biotheor.* **5**, 11–40 (1939).
2. D. G. Kendall, On the generalized "birth-and-death" process. *Ann. Math. Stat.* **19**, 1–15 (1948).
3. S. Nee, R. M. May, P. H. Harvey, The reconstructed evolutionary process. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **344**, 305–311 (1994).
4. H. Morlon, M. D. Potts, J. B. Plotkin, Inferring the dynamics of diversification: A coalescent approach. *PLoS Biol.* **8**, e1000493 (2010).
5. H. Morlon, T. L. Parsons, J. B. Plotkin, Reconciling molecular phylogenies with the fossil record. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 16327–16332 (2011).
6. H. Morlon, Phylogenetic approaches for studying diversification. *Ecol. Lett.* **17**, 508–525 (2014).
7. T. Stadler, D. Kühnert, S. Bonhoeffer, A. J. Drummond, Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci. U.S.A.* **110**, 228–233 (2013).
8. T. B. Quental, C. R. Marshall, Diversity dynamics: Molecular phylogenies need the fossil record. *Trends Ecol. Evol.* **25**, 434–441 (2010).
9. S. Höhna, T. Stadler, F. Ronquist, T. Britton, Inferring speciation and extinction rates under different sampling schemes. *Mol. Biol. Evol.* **28**, 2577–2589 (2011).
10. N. Cusimano, T. Stadler, S. S. Renner, A new method for handling missing species in diversification analysis applicable to randomly or nonrandomly sampled phylogenies. *Syst. Biol.* **61**, 785–792 (2012).
11. M. A. McPeek, J. M. Brown, Clade age and not diversification rate explains species richness among animal taxa. *Am. Nat.* **169**, E97–E106 (2007).
12. F. L. Condamine, J. Rolland, H. Morlon, Assessing the causes of diversification slowdowns: Temperature-dependent and diversity-dependent models receive equivalent support. *Ecol. Lett.* **22**, 1900–1912 (2019).
13. L. Francisco Henao Diaz, L. J. Harmon, M. T. C. Sugawara, E. T. Miller, M. W. Pennell, Macroevolutionary diversification rates show time dependency. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 7403–7408 (2019).
14. T. Stadler, On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *J. Theor. Biol.* **261**, 58–66 (2009).
15. A. Gavryushkina, D. Welch, T. Stadler, A. J. Drummond, Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLOS Comput. Biol.* **10**, e1003919 (2014).
16. S. Louca, M. W. Pennell, Extant timetrees are consistent with a myriad of diversification histories. *Nature* **580**, 502–505 (2020).
17. S. Myers, C. Fefferman, N. Patterson, Can one learn history from the allelic spectrum? *Theor. Popul. Biol.* **73**, 342–348 (2008).
18. A. Bhaskar, Y. S. Song, A. Bhaskar, Y. S. Song, Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Ann. Stat.* **42**, 2469–2493 (2014).
19. H. Morlon, F. Hartig, S. Robin, Prior hypotheses or regularization allow inference of diversification histories from extant timetrees. *bioRxiv* [Preprint] (2020). https://doi.org/10.1101/2020.07.03.185074. Accessed 1 October 2021.
20. M. Pagel. Evolutionary trees can't reveal speciation and extinction rates. *Nature* **580**, 461–462 (2020).
21. A. Helmstetter *et al.*, Pulled Diversification Rates, Lineages-Through-Time Plots, and Modern Macroevolutionary Modeling. *Syst. Biol.* **71**, 758–773 (2022). https://doi.org/10.1101/2021.01.04.424672.
22. L. A. Parry, Evolution: No extinction? No way! *Curr. Biol.* **31**, R907–R909 (2021).
23. J. Kim, E. Mossel, M. Z. Rácz, N. Ross, Can one hear the shape of a population history? *Theor. Popul. Biol.* **100C**, 26–38 (2015).
24. A. Bhaskar, Y. X. Wang, Y. S. Song, Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Res.* **25**, 268–279 (2015).
25. W. S. DeWitt, K. D. Harris, A. P. Ragsdale, K. Harris, Nonparametric coalescent inference of mutation spectrum history and demography. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2013798118 (2021).
26. R. Bouckaert *et al.*, BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019).
27. S. Louca, A. McLaughlin, A. MacPherson, J. B. Joy, M. W. Pennell, Fundamental identifiability limits in molecular epidemiology. *Mol. Biol. Evol.* **38**, 4010–4024 (2021).
28. J. A. Kamm, J. Terhorst, Y. S. Song, Efficient computation of the joint sample frequency spectra for multiple populations. *J. Comput. Graph. Stat.* **26**, 182–194 (2017).
29. A. Polanski, M. Kimmel, New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* **165**, 427–436 (2003).
30. G. J. O. Jameson, Counting zeros of generalised polynomials: Descartes' rule of signs and Laguerre's extensions. *Math. Gaz.* **90**, 223–234 (2006).
31. R. C. Griffiths, S. Tavaré, The age of a mutation in a general coalescent tree. *Commun. Stat. Stoch. Models* **14**, 273–295 (1998).
32. A. Polanski, A. Bobrowski, M. Kimmel, A note on distributions of times to coalescence, under time-dependent population size. *Theor. Popul. Biol.* **63**, 33–40 (2003).
33. P. Paule, M. Schorn, A Mathematica version of Zeilberger's algorithm for proving binomial coefficient identities. *J. Symbolic Comp.* **20**, 673–698 (1995).