A General Birth-Death-Sampling Model for Epidemiology and Macroevolution

Ailene MacPherson[1,2,*], Stilianos Louca[3,4], Angela McLaughlin[5,6], Jeffrey B. Joy[5,6,7] & Matthew W. Pennell[1,*]

[1]*Department of Zoology and Biodiversity Research Centre, University of British Columbia, Vancouver, Canada*
[2]*Department of Ecology and Evolutionary Biology, University of Toronto,Toronto, Canada*
[3]*Department of Biology, University of Oregon, USA*
[4]*Institute of Ecology and Evolution, University of Oregon, USA*
[5]*British Columbia Centre for Excellence in HIV/AIDS, Vancouver, Canada*
[6]*Bioinformatics, University of British Columbia, Vancouver, Canada*
[7]*Department of Medicine, University of British Columbia, Vancouver, Canada*
[*]Correspondence should be addressed to: pennell@zoology.ubc.ca and ailene.macpherson@utoronto.ca

Abstract

Birth-death stochastic processes are the foundation of many phylogenetic models and are widely used to make inferences about epidemiological and macroevolutionary dynamics. There are a large number of birth-death model variants that have been developed; these impose different assumptions about the temporal dynamics of the parameters and about the sampling process. As each of these variants was individually derived, it has been difficult to understand the relationships between them as well as their precise biological and mathematical assumptions. And without a common mathematical foundation, deriving new models is non-trivial. Here we unify these models into a single framework, prove that many previously developed epidemiological and macroevolutionary models are all special cases of a more general model, and illustrate the connections between these variants. To do so, we develop a novel technique for deriving likelihood functions for arbitrarily complex birth-death(-sampling) models that will allow researchers to explore a wider array of scenarios than was previously possible. As an illustration of the utility of our mathematical approach, we use our approach to derive a yet unstudied variant of the birth-death process in which the key rates emerge deterministically from a classic susceptible infected recovered (SIR) epidemiological model.

Keywords: epidemiology; macroevolution; phylogenetics; birth-death processes; statistical inference

Introduction

As a consequence of their rapid mutation rates and large population sizes, many viral pathogens, such as HIV and SARS-CoV-2, accumulate genetic diversity on the timescale of transmission. This genetic diversity can be used to reconstruct the evolutionary relationships between viral variants sampled from different hosts, which in turn can help elucidate the epidemiological dynamics of a pathogen over time. The combined dynamics of viral diversification and the epidemiological process have been termed "phylodynamics" [1, 2]. In the last two decades, there has been a tremendous amount of innovation in phylodynamic models, and the epidemiological inferences from these models increasingly complement those from more conventional surveillance data [2].

Phylodynamic models can be broadly grouped into two classes. The first, based on Kingman's coalescent process [3], has been historically widely used to examine changes in the historical population size of pathogens [4–7]. While useful in some applications, using the coalescent process depends upon the critical assumption that the population size is large relative to the number of samples, such that stochastic variation can be ignored; thus, this approach is inaccurate for reconstructing the dynamics of well-sampled or emerging pathogens [8, 9]. The second class of models, based on the birth-death process [10–13], make no assumption about sparse sampling and fully incorporate stochasticity, and are thus become an increasingly favorable and popular alternative to coalescent models. These birth-death-sampling (BDS) models which are the focus of the present contribution, are also widely utilized in macroevolution to infer speciation and extinction rates over time [14–17] and to estimate divergence times from phylogenetic data [18, 19].

In the context of phylodynamics, BDS models have the additional property that the model parameters, which can be estimated from viral sequence data, explicitly correspond to parameters in classic structured epidemiological models that are often fit to case surveillance data. As the name implies, the BDS process includes three types of events: birth (pathogen transmission between hosts, or speciation in a macroevolutionary context), death (host death or recovery, or extinction in macroevolution), and sampling (including fossil collection and macroevolution). Taken together, these dynamics can be used to describe changes in the basic and effective reproductive ratios (R_0 and R_e, respectively) over time [20, 21] (see Box 1). A common inferential goal is to describe how the frequency of these events, and other derived variables such as R_e, change throughout the course of an epidemic. As we detail below and in the Supplementary Material, there has been an astounding rise in the variety and complexity of BDS model variants to allow for many different assumptions on transmission dynamics and sampling procedures [e.g., 22–24]. While typically not explicitly tied to mechanistic evolutionary processes, there are a similar abundance of macroevolutionary BDS variants that make different assumptions about the trajectory of biodiversity through time [16–19, 25–27].

This flourishing of methods and models has facilitated critical insights into epidemics [28, 29] and the origins of contemporary biodiversity [16, 30]. However, this diversity of models has made it difficult to trace the connections between variants and to understand the precise epidemiological, evolutionary, and sampling processes that are being assumed to be different by each of them. Furthermore, despite their apparent similarity, these models have been derived on a case-by-case basis using different notation and techniques; this creates a substantial barrier for researchers working to develop novel models for new situations. And critically, it is imperative that we understand the general properties of BDS phylogenetic models and the limits of inferences from them [31] and this is difficult to do without considering the full breadth of possible scenarios.

Here we address all of these challenges by deriving the probability of observing a given phylogeny and series of sampling events under a general BDS model. In our derivation we do not assume anything about the functional form (i.e., temporal dynamics) of the various parameters including sampling, the possibility of

sampling ancestors (or not), or how the process was conditioned. While a more general model may be useful for studying the mathematical properties of sets of models [e.g., 31], it is not necessarily useful for statistical inference. However, in deriving a general form of the BDS likelihood, we develop a six-step procedure for deriving the likelihood of any sub-model. This allows us to precisely characterize the implicit and explicit assumptions made by existing sub-models, illustrate the connections between them, and identify their limitations (we work through many examples of this in the Supplementary Material). To further illustrate the advantages of deriving sub-models from the general case, we derive the likelihood of a previously unconsidered model where λ, μ, and ψ emerge deterministically (as opposed to stochastically; see [23]) from the dynamics of an epidemiological compartmental SIR model, which has a close connection with the models discussed here (see Box 1).

Results and Discussion

The general birth-death-sampling model

Model Specification: The BDS stochastic process begins with a single lineage at time t_{or}. We note that this may be considerably older than the age of the most recent common ancestor of an observed sample which is given by t_{MRCA}. While we focus primarily on applications to epidemiology, our approach is agnostic to whether the rates are interpreted as describing pathogen transmission or macroevolutionary diversification.

In the model, transmission/speciation results in the birth of a lineage and occurs at rate $\lambda(\tau)$, where τ ($0 \leq \tau \leq t_{\mathrm{or}}$) is measured in units of time before the present day, such that λ can be time-dependent. We make the common assumption that lineages in the viral phylogeny coalesce exactly at transmission events, thus ignoring the pre-transmission interval inferred in a joint phylogeny of within- and between-host [32]. Throughout, we will use τ as a a general time variable and $t_{\times}$ to denote a specific time of an event $\times$ time units before the present day (see Table S1). Lineage extinction, resulting from host recovery or death in the epidemiological case or the death of all individuals in a population in the macroevolutionary case, occurs at time-dependent rate $\mu(\tau)$. We allow for two distinct types of sampling: lineages are either sampled according to a Poisson process through time $\psi(\tau)$ or binomially at very short intervals, which we term "concerted sampling attempts" (CSAs), where lineages at some specified time t_l are sampled with probability ρ_l ($\vec{\rho}$ denotes a vector of concerted sampling events at different time points). In macroevolutionary studies based only on extant lineages, there is no Poissonian sampling, but a CSA at the present (i.e., $\rho_0 > 0$). In epidemiology, CSAs correspond to large-scale testing efforts (relative to the background rate of testing) in a short amount of time (relative to the rates of viral sequence divergence); for full explanation, see Supplementary Material section S1.2.3. We call these attempts rather than events because if ρ is small or the infection is rare in the population, few or no samples may be obtained. CSAs can also be incorporated into the model by including infinitesimally short spikes in the sampling rate ψ (more precisely, appropriately scaled Dirac distributions). Hence, for simplicity, in the main text we focus on the seemingly simpler case of pure Poissonian sampling through time except at present-day, where we allow for a CSA to facilitate comparisons with macroevolutionary models; the resulting formulas can then be used to derive a likelihood formula for the case where past CSAs are included (Supplement S1.2.3).

In the epidemiological case, sampling may be concurrent (or not) with host treatment or behavioural changes resulting in the effective extinction of the viral lineage. Hence, we assume that sampling results in the immediate extinction of the lineage with probability $r(\tau)$. Similarly, in the case of past CSAs we must include the probability, r_l, that sampled hosts are removed from the infectious pool during the CSA at time

t_l. Poissonian sampling without the removal of lineages ($r(\tau) < 1$) can be employed in the macroevolutionary case to explicitly model the collection of samples from the fossil record (e.g., the fossilized birth-death process [19]).

For our derivation, we make no assumption about the temporal dynamics of λ, μ, ψ, or r; each may be constant, or vary according to any arbitrary function of time given that it is biologically valid (i.e., non-negative and between 0 and 1 in the case of r). We make the standard assumption that at any given time any given lineage experiences a birth, death or sampling event independently of (and with the same probabilities as) all other lineages. We revisit this assumption in Box 1 where we discus how the implicit assumptions of the BDS process are well summarized by the diversification model's relationship to the SIR epidemiological model. Our resulting general time-variable BDS process can be fully defined by the parameter set $\Theta_{\mathrm{BDS}} = \{\lambda(\tau), \mu(\tau), \psi(\tau), r(\tau), \vec{\rho}\}$.

In order to make inference about the model parameters, we need to calculate the likelihood, $\mathcal{L}$, that an observed phylogeny, $\mathcal{T}$, is the result of a given BDS process. With respect to the BDS process there are two ways to represent the information contained in the phylogeny $\mathcal{T}$, both of which have been used in the literature, which we call the "edge" and "critical time" representations, respectively. We begin by deriving the likelihood in terms of the edge representation and later demonstrate how to reformulate the likelihood in terms of critical times. In the edge representation, the phylogeny is summarized as a set of edges in the mathematical graph that makes up the phylogeny, numbered 1-11 in Figure B1, and the types of events that occurred at each node. We define $g_e(\tau)$ as the probability that the edge e which begins at time s_e and ends at time t_e gives rise to the subsequently observed phylogeny between time τ, $(s_e < \tau < t_e)$ and the present day. The likelihood of the tree then, is by definition $g_{\mathrm{stem}}(t_{\mathrm{or}})$: the probability density the stem lineage ($\mathrm{stem} = 1$ in Figure B1) gives rise to the observed phylogeny from the origin, t_{or}, to the present day. Although it is initially most intuitive to derive the likelihood in terms of the edge representation, as we show below, it is then straightforward to derive the critical times formulation which results in mathematical simplification.

Deriving the Initial Value Problem (IVP) for $g_e(\tau)$: We derive the IVP for the likelihood density $g_e(\tau)$ using an approach first developed by [11]. For a small time $\Delta\tau$ the recursion equation for the change in the likelihood density is given by the following expression.

$$\begin{aligned} g_e(\tau + \Delta\tau) = &\underbrace{(1 - \lambda(\tau)\Delta\tau)(1 - \mu(\tau)\Delta\tau)(1 - \psi(\tau)\Delta\tau) \times g_e(\tau)}_{\text{nothing happens}} \\ &+ \underbrace{\lambda(\tau)\Delta\tau(1 - \mu(\tau)\Delta\tau)(1 - \psi(\tau)\Delta\tau) \times 2g_e(\tau)E(\tau)}_{\text{birth event}} \\ &+ \underbrace{\mu(\tau)\Delta\tau(1 - \lambda(\tau)\Delta\tau)(1 - \psi(\tau)\Delta\tau) \times 0}_{\text{death event}} \\ &+ \underbrace{\psi(\tau)\Delta\tau(1 - \lambda(\tau)\Delta\tau)(1 - \mu(\tau)\Delta\tau) \times 0}_{\text{sampling event}}. \end{aligned} \tag{1}$$

Here, $E(\tau)$ is the probability that a lineage alive at time τ leaves no sampled descendants at the present day. We will examine this probability in more detail below. Assuming $\Delta\tau$ is small, we can approximate the above recursion equation as the following difference equation.

$$\Delta g_e(\tau) \approx -(\lambda(\tau) + \mu(\tau) + \psi(\tau))\Delta\tau g_e(\tau) + 2\lambda(\tau)g_e(\tau)E(\tau)\Delta\tau + \mathcal{O}(\Delta\tau^2). \tag{2}$$

By the definition of the derivative we have:

$$\frac{dg_e(\tau)}{d\tau} = -(\lambda(\tau) + \mu(\tau) + \psi(\tau))g_e(\tau) + 2\lambda(\tau)g_e(\tau)E(\tau). \tag{3}$$

Equation (3) is known as the Kolmogorov backward equation of the BDS process [33–35]. Beginning at time s_e, the initial condition of g_e depends on which event occurred at the beginning of edge e.

$$g_e(s_e) = \begin{cases} \lambda(s_e)g_{e1}(s_e)g_{e2}(s_e) & \text{birth event giving rise to edges e1 and e2} \\ (1-r(s_e))\psi(s_e)g_{e1}(s_e) & \text{ancestral sampling event} \\ \psi(s_e)r(s_e) + \psi(s_e)(1-r(s_e))E(s_e) & \text{terminal sampling event} \\ \rho_0 & s_e = 0\text{, extant sample} \end{cases} \tag{4}$$

Together Equations (3) and (4) define the initial value problem for $g_e(\tau)$ as a function of the probability $E(\tau)$.

Because the likelihood density g_e is the solution to a linear differential equation with initial condition at time s_e, we can express its solution as follows:

$$g_e(\tau) = \Psi(s_e, \tau)g_e(s_e), \tag{5}$$

where the auxiliary function, Ψ, is given by:

$$\Psi(s_e, \tau) = \exp\left[\int_{s_e}^{\tau} 2\lambda(x)E(x) - (\lambda(x) + \mu(x) + \psi(x))\, dx\right]. \tag{6}$$

This function, $\Psi(s, t)$, maps the value of g_e at time s to its value at t, and hence is known as the probability "flow" of the Kolmogorov backward equation [35].

Deriving the IVP for $E(\tau)$: We derive the IVP for $E(\tau)$ in a similar manner as above, beginning with a difference equation.

$$\begin{aligned} E(\tau + \Delta\tau) = &\underbrace{(1-\lambda(\tau)\Delta\tau)(1-\mu(\tau)\Delta\tau)(1-\psi(\tau)\Delta\tau) \times E(\tau)}_{\text{nothing happens}} \\ &+ \underbrace{\lambda(\tau)\Delta\tau(1-\mu(\tau)\Delta\tau)(1-\psi(\tau)\Delta\tau) \times E(\tau)^2}_{\text{birth event}} \\ &+ \underbrace{\mu(\tau)\Delta\tau(1-\lambda(\tau)\Delta\tau)(1-\psi(\tau)\Delta\tau) \times 1}_{\text{death event}} \\ &+ \underbrace{\psi(\tau)\Delta\tau(1-\lambda(\tau)\Delta\tau)(1-\mu(\tau)\Delta\tau) \times 0}_{\text{sampling event}}. \end{aligned} \tag{7}$$

By the definition of a derivative we have:

$$\begin{aligned} \frac{dE(\tau)}{d\tau} =& -(\lambda(\tau) + \mu(\tau) + \psi(\tau))E(\tau) + \lambda(\tau)E(\tau)^2 + \mu(\tau), \\ E(0) =& 1 - \rho_0, \end{aligned} \tag{8}$$

where ρ_0 is the probability a lineage is sampled at the present day. The initial condition at time 0 is therefore the probability that a lineage alive at the present day is not sampled. Given an analytical or numerical general solution to $E(\tau)$, we can find the likelihood by evaluating $g_{stem}(t_{or})$, as follows.

Deriving the expression for $g_{stem}(t_{or})$: Given the linear nature of the differential equation for $g_e(\tau)$ and hence the representation in Equation (5)), the likelihood $g_{\text{stem}}(\tau)$ is given by the product over all the initial

conditions times the product over the probability flow for each edge.

$$g_{stem}(t_{\text{or}}) = \underbrace{\rho_0^{N_0}}_{\substack{\text{extant}\\\text{tips}}} \underbrace{\prod_{i=1}^{I} \lambda(x_i)}_{\text{births}} \underbrace{\prod_{j=1}^{n} [\psi(y_j)(1-r(y_j))E(y_j) + \psi(y_j)r(y_j)]}_{\text{extinct tips}} \times \underbrace{\prod_{k=1}^{m} \psi(z_k)(1-r(z_k))}_{\text{ancestral samples}} \underbrace{\prod_{e\in\mathcal{T}} \Psi(s_e, t_e)}_{\text{edges}} . \tag{9}$$

Representing $g_{stem}(t_{or})$ in terms of critical times: Equation (9) can be further simplified by removing the need to enumerate over all the edges of the phylogeny (the last term of Equation (9)) and writing $\mathcal{L}$ in terms of the tree's critical times (horizontal lines in figure B1). The critical times of the tree are made up of three vectors, $\vec{x}$, $\vec{y}$, and $\vec{z}$, as well as the time of origin t_{or}. The vector $\vec{x}$ gives the time of each birth event in the phylogeny and has length $I = N_0 - n - 1$ where N_0 is the number of lineages sampled at the present day and n is the number of terminal samples. Unless noted otherwise the elements of vector $\vec{x}$ are listed in decreasing order, such that $x_1 > x_2 > ...x_I$ and hence x_1 is the time of the most recent common ancestor t_{MRCA}. The vector $\vec{y}$ gives the timing of each terminal sample and hence has length n whereas vector $\vec{z}$ gives the timing of each ancestral sample and has length m. With respect to the BDS likelihood then the sampled tree is summarized by $\mathcal{T} = \{\vec{x}, \vec{y}, \vec{z}, t_{or}\}$. We note that the critical times only contain the same information as the edges as a result of the assumptions of the BDS process but are not generally equivalent representations of $\mathcal{T}$.

As a result of the linear nature of $g_e(\tau)$ it is straightforward to rewrite the likelihood in Equation (9) in terms of the critical-time representation of the sampled tree. Defining

$$\Phi(t) = \Psi(0,t) = \exp\left[\int_0^{\tau} 2\lambda(x)E(x) - (\lambda(x) + \mu(x) + \psi(x))\, dx\right], \tag{10}$$

the probability flow Ψ can be rewritten as the following ratio:

$$\Psi(s,t) = \frac{\Psi(0,\tau)}{\Psi(0,s)} = \frac{\Phi(s)}{\Phi(t)}. \tag{11}$$

This relationship allows us to rewrite the likelihood by expressing the product over the edges as two separate products, one over the start of each edge and the other over the end of each edge. Edges begin (value of t_e) at either: 1) the tree origin, 2) a birth event resulting to two lineages, or 3) an ancestral sampling event. Edges end (values of s_e) at either: 1) a birth event, 2) an ancestral sampling event, 3) a terminal sampling event, or 4) the present day. Hence we have:

$$g_{stem}(t_{\text{or}}) = \underbrace{\Phi(t_{\text{or}})}_{\text{root}} \times \underbrace{\left(\frac{\rho_0}{\cancel{\Phi(0)}}\right)^{N_0}}_{\text{extant tips}} \times \underbrace{\prod_{i=1}^{I} \lambda(x_i) \frac{\Phi(x_i)^{\cancel{2}}}{\cancel{\Phi(x_i)}}}_{\text{births}} \times \underbrace{\prod_{j=1}^{n} \frac{\psi(y_j)}{\Phi(y_j)} [(1-r(y_j))E(y_j) + r(y_j)]}_{\text{extinct tips}} \times \underbrace{\prod_{k=1}^{m} \frac{\cancel{\Phi(z_k)}}{\cancel{\Phi(z_k)}} \psi(z_k)(1-r(z_k))}_{\text{ancestral samples}} . \tag{12}$$

Note $\Phi(0) = 1$.

Conditioning the likelihood: While Equation (12) is equal to the basic likelihood of the phylogeny $\mathcal{T}$, it is often appropriate to condition the tree likelihood on the tree exhibiting some property, for example the condition there being at least sampled lineage. Imposing a condition on the likelihood is done by multiplying by a factor $\mathcal{S}$. Various conditioning schemes are considered in section S1.4 and listed in Table S3. The resulting likelihood of the general BDS model is:

$$\boxed{\begin{aligned}\mathcal{L}\left(\vec{x},\vec{y},\vec{z},N_0|\Theta_{\mathrm{BDS}},\mathcal{S}\right) =& \mathcal{S}\rho_0^{N_0}\Phi(t_{\mathrm{or}})\prod_{i=1}^{I}\lambda(x_i)\Phi(x_i)\\ &\times\prod_{j=1}^{n}\frac{\psi(y_j)}{\Phi(y_j)}\left[(1-r(y_j))E(y_j)+r(y_j)\right]\prod_{k=1}^{m}\psi(z_k)\left(1-r(z_k)\right)\end{aligned}} \tag{13}$$

Many existing models are special cases of our general BDS

A large variety of previously published BDS models in epidemiology and macroevolution are simply special cases of the general model presented here (for a summary of the models we investigated see Table S2; proofs in Supplemental Material). Indeed, we can obtain the likelihood of these models by adding mathematical constraints (i.e., simplifying assumptions) to the terms in Equation (13). Our work thus not only provides a consistent notation for unifying a multitude of seemingly disparate models, it also provides a concrete and numerically straightforward recipe for computing their likelihood functions. We have implemented the general form of the BDS framework in the R package `castor` [36], including routines for maximum-likelihood fitting of the BDS models with arbitrary functional forms of the parameters given a phylogeny and routines for simulating phylogenies under the general BDS models (functions `fit_hbds_model_on_grid`, `fit_hbds_model_parametric` and `generate_tree_hbds`).

Figure 1 summarizes the simplifying assumptions that underlie common previously published BDS models; these assumptions generally fall into four categories: 1) assumptions about the functional form of birth, death, and sampling rates over time, 2) assumptions pertaining to the sampling of lineages, 3) the presence of mass-extinction events, and 4) the nature of the tree-conditioning as given by $\mathcal{S}$. Here we provide a brief overview of the type of previously-invoked constraints which are consistent (or not) with our generalized BDS; for full details on each specific case, we refer readers to the Supplementary Material.

In regards to rate assumptions, many early BDS models [12, 13, 20] assumed that the birth, death, and sampling rates remained constant over time. This is mathematically and computationally convenient since an analytical solution can easily be obtained for $E(\tau)$. In the epidemiological case, holding λ constant, however, implies that the number of susceptible hosts is effectively constant throughout the epidemic and/or that the population does not change its behavior over time (an unrealistic assumption, e.g. in the face of seasonal changes or changes in response to the disease itself). As such, this assumption is only really valid for small time periods or the early stages of an epidemic. This is useful for estimating the basic reproductive number, R_0, of the SIR model (Box 1) but not for the effective reproductive number R_e at later time points [20].

A similarly tractable, but more epidemiologically relevant, model is known as the "birth-death-skyline" variant [21, 24], in which rates are piecewise-constant functions through time (like the constant rate model, there is also an analytical way to calculate the likelihood of this model; see Supplementary Material S1.1.2). The BDS skyline model has been implemented under a variety of additional assumptions in the Bayesian phylogenetics software BEAST [37]. The BDS skyline model has also been extended by Kuhnert et al. [23] to infer the the parameters of an underlying stochastic SIR model. In this case the diversification model parameters Θ_{BDS} are random variables that emerge from stochastic realizations of the epidemiological model

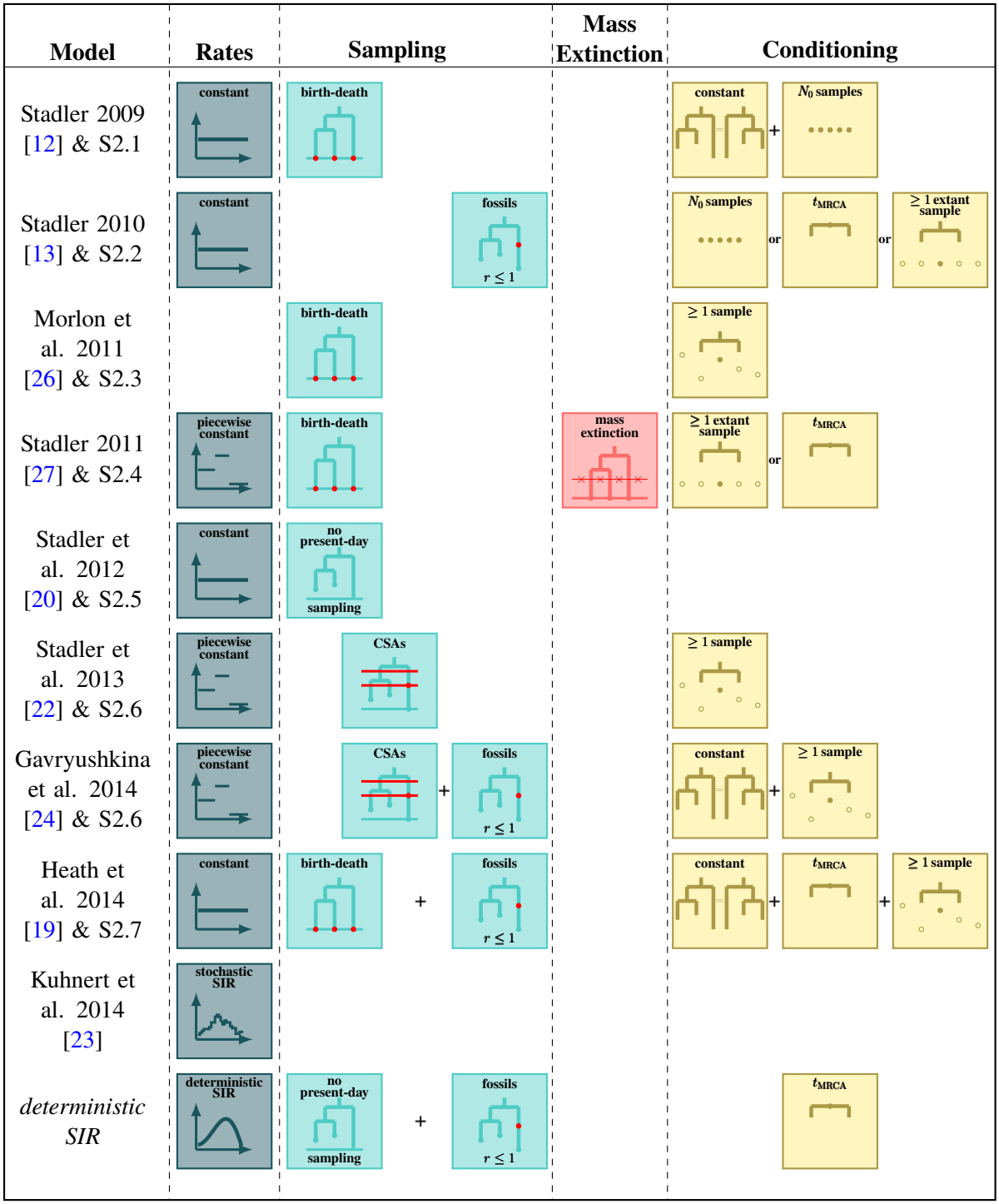

Figure 1: **Sub-model assumptions.** Rate, sampling, mass extinction, and conditioning assumptions of existing sub-models of the general time-variable BDS process. The key points are that i) each of the previously developed models we considered can be obtained by adding specific combinations of constraints to the various parameters of the general BDS model; and ii) that there are many plausible, and potentially biologically informative combinations of constraints that have not been considered by researchers in epidemiology or macroevolution.

given by Θ_{SIR}, see Equation (B1). Finally, the birth-death skyline model with piecewise constant rates can also be applied in the macroevolutionary case when no sampling occurs through time, $\psi(\tau) = 0$ [27].

In addition to imposing constraints on the temporal variation in the rates, previously derived sub-models have considered a variety of different assumptions about the nature of the sampling process. Most notably, in macroevolutionary studies, sampling of molecular data typically occurs only at the present day [12, 26, 27] whereas past Poissonian sampling can be introduced to include the sampling of fossil data [19]. In epidemiology, concerted sampling at the present day is likely biologically unrealistic (e.g., [20] and the deterministic SIR model considered below), though in some implementations of the models, such a sampling scheme has been imposed. Concerted sampling attempts prior to the present day as well as mass extinction events can be incorporated, as mentioned previously, via the inclusion of Dirac distributions in the sampling and death rates, respectively. Finally, previous models include multiplying the likelihood by a factor $\mathcal{S}$ in order to condition on a particular observation (e.g. observing at least one lineage or exactly N_0 lineages), enumerate of indistinguishable trees [12, 24, 38], or to reflect known uncertainties. The "fossilized-birth-death" likelihood derived by Heath et al. [19] for example, includes a factor that reflects the uncertainty in the attachment and placement of fossils on the macroevolutionary tree. The Supplementary Material demonstrates how these sub-models can be re-derived by either imposing the necessary constraints on the general likelihood formula given in Equation (13) or, alternatively, by starting from the combinations of assumptions and using the six-step procedure outlined above (and as illustrated in the example below).

Deriving a new BDS variant: a deterministic SIR model

As we mentioned above and proved in the Supplementary Materials, many previously derived BDS variants are sub-models of a more general time-variable BDS that we derive here. Figure 1 illustrates how one can add (and combine) additional mathematical constraints or biological assumptions to derive a specific variant. However, the range of BDS models published so far is not exhaustive and there are a multitude of alternative ways that one could constrain the functional forms of the parameters. As we emphasized before, it is straightforward to derive the likelihood of any of these variants using the approach we present here. As an illustrative example, we derive the likelihood of a novel model motivated by the close connection between the BDS process and SIR models in epidemiology (Box 1). As noted above, Kuhnert *et al.* [23] previously derived a model in which the rates of the BDS process are random variables that emerge stochastically from the dynamics of the underlying SIR model. Allowing for this stochasticity is critically important at early stages of an epidemic when viral population sizes are small [8, 23]; however, later in an epidemic when infections are widespread, it may be more tractable to fit a model where the BDS rates are deterministic. A detailed analysis of when a stochastic or deterministic model might be preferred are beyond the scope of the present contribution (see [39]) — we re-emphasize that our primary purpose is to illustrate that expanding the range of available models is straightforward given our general formulation.

The Epidemiological Model: Here let t denote time measured forward in time and τ as time measured backward from the present, with the two related via $t = t_{or} - \tau$. The epidemic originates at time $t = 0$ and $\tau = t_{or}$ and the present day is given by $t = t_{or}$ and $\tau = 0$. In this case we use the SIR model as defined in Box 1 assuming the epidemiological rates (e.g. $\beta\,\gamma$, $\alpha\,\psi$ and σ in Figure B1) are time independent. In this

case the deterministic compartmental model is given by the following initial value problem.

$$\begin{aligned}
\frac{dS(t)}{dt} =& bN(t)\left(1-\frac{N(t)}{\kappa}\right) - \delta S(t) - \beta S(t)I(t) + \sigma R(t) \quad S(0) = (1-f_I)N(0) \\
\frac{dI(t)}{dt} =& \beta S(t)I(t) - (\delta+\alpha+\gamma+\psi r)\, I(t) \quad I(0) = f_I N(0) \\
\frac{dR(t)}{dt} =& (\gamma+\psi r)\, I(t) - (\delta+\sigma)\, R(t) \quad R(0) = 0
\end{aligned} \tag{14}$$

where $N(t) = S(t) + I(t) + R(t)$ is the total number of hosts. Here we use a model of logistic growth for host birth B in which the pathogen has no suppressive effects on reproduction. The initial conditions are chosen such that epidemic originates with a small fraction f_I of infected hosts with no pre-existing immunity in the population. System (14) has no known general analytical solution but can be solved numerically. We denote this numerical solution for the number of hosts in each compartment with $\tilde{S}(t)$, $\tilde{I}(t)$ and $\tilde{R}(t)$.

Step 1: Model Specification — As discussed in box 1, SIR compartmental models can be used to constrain BDS rates by setting $\lambda(\tau) = \beta\tilde{S}(t_{or}-\tau)$ and $\mu(\tau) = \gamma+\delta+\alpha$, which under the present assumptions is a constant. The sampling rate is also assumed constant $\psi(\tau) = \psi$. We will assume that all samples are acquired through Poissonian sampling at a constant rate, hence we include neither CSAs at in the past nor at the present-day. Upon sampling we will assume that all lineages are removed with a constant probability r. Finally, we will condition the likelihood on the observation of at least one sample since the time of the most recent common ancestor. To avoid extending the inference to points very early in the spread of the infection we will condition on observing at least one sampled lineage since the time of most recent common ancestor, $\mathcal{S}_6$.

Step 2: IVP for $g_e(\tau)$ — The initial value problem for $g_e(\tau)$ is straightforward to derive:

$$\begin{aligned}
\frac{dg_e(\tau)}{d\tau} &= -\left(\beta\tilde{S}(t_{or}-\tau)+\gamma+\delta+\alpha+\psi\right) g_e(\tau) + 2\beta\tilde{S}(t_{or}-\tau)g_e(\tau)E(\tau) \\
g_e(s_e) &= \begin{cases} \beta\tilde{S}(t_{or}-s_e)g_{e1}(s_e)g_{e2}(s_e) & \text{birth event giving rise to edges e1 and e2} \\ (1-r)\psi(s_e)g_e(s_e) & \text{ancestral sampling event} \\ \psi r + \psi(1-r)E(s_e) & \text{terminal sampling event} \end{cases}
\end{aligned}$$

Step 3: IVP for $E(\tau)$ — Similarly we have:

$$\frac{dE(\tau)}{d\tau} = -\left(\beta\tilde{S}(t_{or}-\tau)+\gamma+\delta+\alpha+\psi\right) E(\tau) + \beta\tilde{S}(t_{or}-\tau)E(\tau)^2 + \gamma+\delta+\alpha \quad \text{where } E(0) = 1$$

Step 4: Expression for $g_{stem}(t_{or})$. The expression for $g_{stem}(t_{or})$ in the case of no present-day sampling can be obtained from 9 and setting $N_0 = 0$.

$$g_{stem}(t_{\text{or}}) = \underbrace{\prod_{j=1}^{n}\left(\psi(1-r)E(y_j)+\psi r\right)}_{\text{extinct tips}} \underbrace{(\psi(1-r))^m}_{\text{ancestral samples}} \underbrace{\prod_{i=1}^{I}\beta\tilde{S}(t_{or}-x_i)}_{\text{births}} \underbrace{\prod_{e\in\mathcal{T}}\Psi(s_e,t_e)}_{\text{edges}},$$

Step 5: $g_{stem}(t_{or})$ in terms of the critical times — We can then transform the solution into the critical time

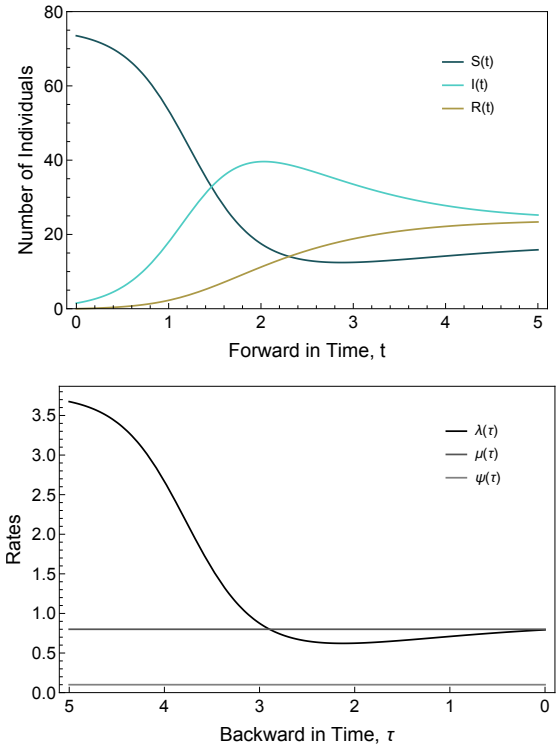

Figure 2: **Epidemiological model dynamics and corresponding diversification rates.** Top panel: deterministic epidemiological dynamics of system (14) as given by the number of susceptible $S(t)$, infected$I(t)$, and recovered $R(t)$ hosts as a function of time t measured forward in time. Bottom Panel: Corresponding diversification model as given by the birth $\lambda(\tau)$ and death $\mu(\tau)$ rates as a function of time τ as measured backwards in time. Parameters: $b = 1$, $\delta = 0.25$, $\kappa = 100$, $\beta = 0.05$, $\sigma = 0.1$, $\alpha = 0.25$, $\psi = 0.1$, $\gamma = 0.3$, $r = 0.5$, $f_I = 0.02$, and $t_{or} = 5$.

representation given the relation $\Psi(t,s) = \Phi(t)/\Phi(s)$.

$$\Psi(\tau) = \exp\left[\int_0^\tau \beta\tilde{S}(t_{or}-x)(2E(x)-1) - (\gamma+\delta+\alpha+\psi)\,dx\right]$$

$$g_{stem}(t_{or}) = \underbrace{\Phi(t_{\text{or}})}_{\text{root}} \underbrace{\prod_{i=1}^{n+m-1} \beta\tilde{S}(t_{or}-x_i)\Phi(x_i)}_{\text{births}} \underbrace{\prod_{j=1}^{n} \frac{\psi}{\Phi(y_j)}\left[(1-r)E(y_j)+r\right]}_{\text{extinct tips}} \underbrace{(\psi(1-r))^m}_{\text{ancestral samples}}.$$

Step 6: Conditioning the likelihood — Finally, we condition the likelihood on observing at least one lineage given the TMRCA, $\mathcal{S}_5 = \frac{\Phi(\tau_{x_1})}{\Phi(\tau_{\text{or}})(1-E(x_1))^2}$,

$$\boxed{\mathcal{L}(\mathcal{T}|\Theta_{SIR}) = \frac{\Phi(\tau_{x_1})}{(1-E(x_1))^2} \prod_{i=1}^{n+m-1} \beta\tilde{S}(t_{or}-x_i)\Phi(x_i) \prod_{j=1}^{n} \frac{\psi}{\Phi(y_j)}\left[(1-r)E(y_j)+r\right](\psi(1-r))^m} \tag{15}$$

Equation (15) can then be used to estimate the underlying parameters of the SIR model from a phylogeny. Again, while the relative costs and benefits of using the deterministic versus stochastic formulation [23] of the model are beyond our scope, this example serves mainly to illustrate that the likelihood functions of complex and biologically interesting new model variants are straightforward to derive using our mathematical framework.

Conclusion

Here we derive a phylogenetic birth-death-sampling model in a more general form than previously attempted, making as few assumptions about the processes that generated the data as possible. While drawing inferences from data will require making additional assumptions and applying mathematical constraints to the parameters (but see [40]), a general unifying model can clarify the connections between various model variants, provide a framework for developing new variants tailored specifically to each situation, and provide a framework for understanding how results depend on model assumptions [31, 41, 42]. From a methodological perspective, our technique for deriving the likelihood of BDS models substantially lowers the barrier for developing and exploring new types of models in a way that Maddison *et al.* [11] did for birth-death models without heterogeneous sampling. In fact, the fully general and yet numerically tractable likelihood formula presented here allowed us to implement computational methods for simulating and fitting BDS models with arbitrary functional forms for λ, μ, ψ and r [36].

There are also a multitude of previously developed models that we did not explore here, most notably multi-type models where different lineages have different rates; such multi-type models have become very popular in macroevolution and phylodynamics [8, 11, 26, 43, 44]. A natural extension to the present work would be to extend our approach to these cases, which in most cases we suspect is possible. Similarly, we have not explored whether the variants of the BDS process used to reconstruct rates with fossil occurrences [14, 45, 46] and think this would be worth investigating. And perhaps most importantly, given the recent discovery of widespread non-identifiability for models fit to extant-only time trees [31], it is critical to investigate and fully characterize the identifiability of the general BDS process, i.e., when all rates can in principle vary freely over time. The general formulation we have developed here sets the stage for such investigations.

Box 1: The connection between BDS and SIR models

The general BDS model is intimately related to the SIR compartmental model used in classic theoretical epidemiology. This connection illustrates the explicit and implicit assumptions of the general BDS model and its sub models. Here we define the SIR epidemiological model, discuss how it can inform and be informed by these diversification models, and examine the shared assumptions of the two frameworks.

The SIR model:

The SIR model is host-centric, partitioning the population via infection status into susceptible (S), infected (I), and recovered (R) hosts. Infection of susceptible hosts occurs at a per-capita rate βI. Infected hosts may recover (at rate γ), die of virulent cases (at rate α), or be sampled (at rate ψ). The cumulative number of sampled hosts is represented in the SIR model shown in Figure B1 by I^*. Upon sampling, infected hosts may be treated and hence effectively recover with probability r. Hosts that have recovered from infection exhibit temporary immunity to future infection which wanes at rate σ. The special case of the SIR model with no immunity (the SIS model) is obtained in the limit as $\sigma \to \infty$. In addition to these epidemiological processes, the SIR model includes demographic processes, such as host birth (rate B) and death from natural causes (rate δ). While not shown explicitly in the figure, these epidemiological and demographic rates may change over time as a result of host behavioural change, pharmaceutical and non-pharmaceutical interventions, or host/pathogen evolution. In the deterministic (infinite population size) limit, the differential equations for the resulting epidemiological dynamics are given by Equation (14).

The BDS Model:

The BDS model is pathogen-centric, following the number of sampled and unsampled viral lineages over time, analogous to the I and I^* classes of the SIR model. A key element of general BDS model is that birth and death rates may vary over time. This time dependence may be either continuous (e.g. [26, 47]) or discrete (e.g. [21, 23, 24, 27]) Although arbitrarily time-dependent, the birth, death, and sampling rates in the general BDS model are assumed to be diversity-independent, analogous to the assumption of density-dependent transmission (pseudo mass action) in the SIR model [48]. While some forms of diversity-dependence in diversification rates may be incorporated implicitly [49], explicit diversity-dependence (e.g. [50]) goes beyond the scope of the BDS models considered here.

The general BDS model assumes all viral lineages are exchangeable - this has several implications. First, only a single pathogen type exists. Multi-type models (e.g. [22, 43, 44, 51]) are not included in the GBDS framework. Second, transmission is independent of lineage age. In the macroevolutionary case, such age-dependence has been suggested to reflect niche differentiation in novel species [52] and in the epidemiological case may reflect adaptation towards increased transmissibility following a host species-jumping event. Third, lineage exchangeablity is reflected in the absence of an exposed (E) class in the SIR model in which hosts can, for example, transmit infections but not be sampled or vice versa. Finally, the general BDS model assumes all lineages are sampled at random and does not include sub-models with non-random representation of lineages (e.g. [20]).

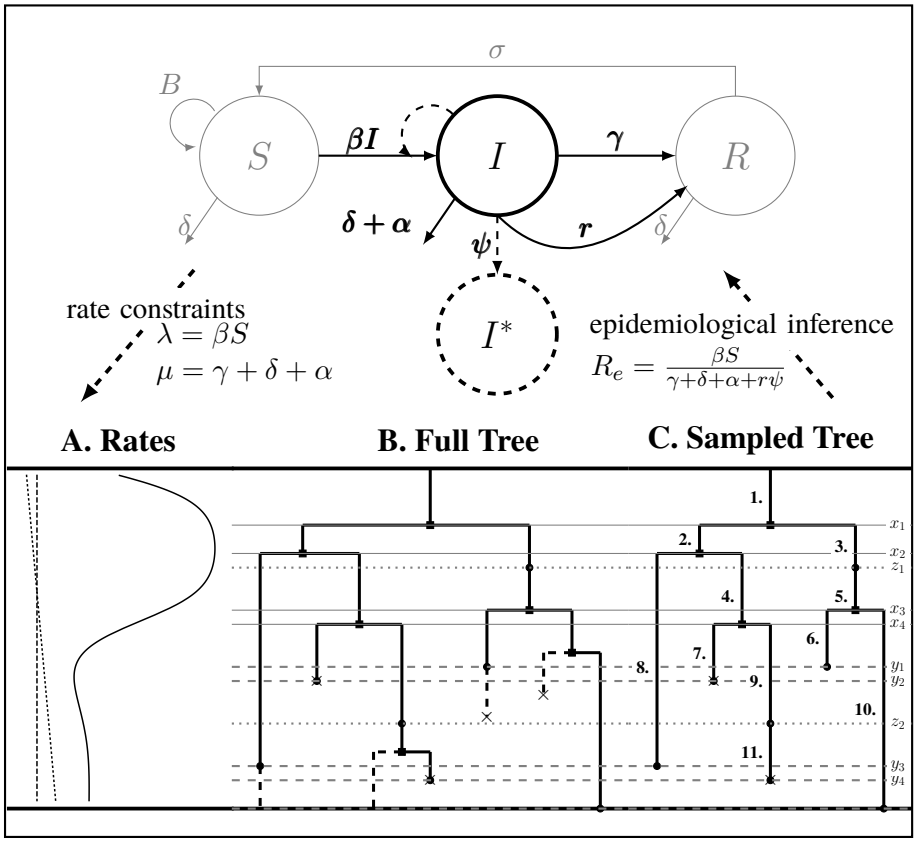

Figure B1: Top: The SIR epidemiological model. Black (gray) lines and classes represent rates and variables followed (in)directly by the BDS model. The SIR model can be used to constrain the rates of the BDS model (A.). Simulated forward in time, the result of the BDS stochastic processes is a *full tree* (B.) giving the complete genealogy of the viral population. Pruning away extinct and unsampled lineages produces the *sampled tree* (C.). Arising from a BDS process, this sampled tree can be summarized in two ways. First by the set of edges (labeled 1-11) or as a set of critical times (horizontal lines) including: 1) the time of birth events (solid, x_i) 2) terminal sampling times (dashed, y_j), and 3) ancestral sampling times (dotted, z_k). Given the inferred rates from a reconstructed sampled tree, these rates can be used to estimate characteristic parameters of the SIR model, for example the basic or effective reproductive number.

Model Connections

Given their shared model assumptions, the general BDS model can be constrained explicitly to reflect an underlying SIR epidemic by setting the viral birth rate equal to the per-capita transmission rate of the infectious class (e.g., $\lambda(\tau) = \beta S(\tau)$) and the viral death rate to the infectious recovery or removal rate $\mu(\tau) = \gamma + \delta + \alpha$, whereas the sampling rate $\psi(\tau)$ is identical across models. While constraining the birth, death, and sampling rates in this manner can be used to parameterize compartmental models (e.g., [23]) doing so is an approximation assuming independence between the exact timing of transmission, recovery or removal from population, and sampling events in the SIR model and birth, death, and sampling events in the diversification model. The resulting tree likelihood in terms of the compartmental model is given by:

$$Pr(\mathcal{T}|\Theta_{SIR}) = \underbrace{Pr(\mathcal{T}|\Theta_{BDS})}_{\text{BDS likelihood}} \underbrace{P(\Theta_{BDS}|\Theta_{SIR})}_{\text{SIR process}}. \tag{B1}$$

While they are not sub-models of the general BDS process, likelihood models have been developed that capture the full non-independence of viral diversification and epidemiological dynamics for the SIR model specifically [53] and in compartmental models in general [54]. The connection between the BDS process and SIR epidemiological models can also be used after the diversification rates are inferred to estimate the basic and effective reproductive rates [20, 21]. Specifically, the effective reproductive rate at time τ before the present day is given by $R_e(\tau) = \frac{\lambda(\tau)}{\mu(\tau)+r(\tau)\psi(\tau)}$.

References

[1] Grenfell BT et al. (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. *Science (New York, N.Y.)* 303(5656):327–332.

[2] Volz EM (2012) Complex Population Dynamics and the Coalescent Under Neutrality. *Genetics* 190(1):187–201.

[3] Kingman JFC (1982) On the Genealogy of Large Populations. *Journal of Applied Probability* 19:27–43.

[4] Pybus OG, Rambaut A, Harvey PH (2000) An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies. *Genetics* 155(3):1429–1437.

[5] Strimmer K, Pybus OG (2001) Exploring the Demographic History of DNA Sequences Using the Generalized Skyline Plot. *Molecular Biology and Evolution* 18(12):2298–2305.

[6] Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution* 22(5):1185–1192.

[7] Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW (2009) Phylodynamics of Infectious Disease Epidemics. *Genetics* 183(4):1421–1430.

[8] Stadler T et al. (2015) How well can the exponential-growth coalescent approximate constant-rate birth-death population dynamics? *Proceedings. Biological Sciences* 282(1806):20150420.

[9] Boskova V, Bonhoeffer S, Stadler T (2014) Inference of Epidemiological Dynamics Based on Simulated Phylogenies Using Birth-Death and Coalescent Models. *PLoS Computational Biology* 10(11).

[10] Kendall DG (1948) On the Generalized "Birth-and-Death" Process. *Annals of Mathematical Statistics* 19(1):1–15.

[11] Maddison WP, Midford PE, Otto SP (2007) Estimating a Binary Character's Effect on Speciation and Extinction. *Systematic Biology* 56(5):701–710.

[12] Stadler T (2009) On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology* 261(1):58–66.

[13] Stadler T (2010) Sampling-through-time in birth-death trees. *Journal of Theoretical Biology* 267(3):396–404.

[14] Raup DM (1985) Mathematical models of cladogenesis. *Paleobiology* 11(1):42–52.

[15] Nee S, May RM, Harvey PH (1994) The reconstructed evolutionary process. *Phil. Trans. R. Soc. B* p. 7.

[16] Morlon H (2014) Phylogenetic approaches for studying diversification. *Ecology Letters* 17(4):508–525. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.12251.

[17] Louca S (2020) Simulating trees with millions of species. *Bioinformatics*.

[18] Gernhard T (2008) The conditioned reconstructed process. *Journal of Theoretical Biology* 253(4):769–778.

[19] Heath TA, Huelsenbeck JP, Stadler T (2014) The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences* 111(29):E2957–E2966.

[20] Stadler T, Kouyos RD, von Wyl V, Yearly S, Böni J (2012) Estimating the Basic Reproductive Number from Viral Sequence Data. *Mol. Biol. Evol.*

[21] Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ (2013) Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *PNAS* 110(1):228–233.

[22] Stadler T, Bonhoeffer S (2013) Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 368(1614):20120198.

[23] Kühnert D, Stadler T, Vaughan TG, Drummond AJ (2014) Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. *Journal of The Royal Society Interface* 11(94):20131106.

[24] Gavryushkina A, Welch D, Stadler T, Drummond AJ (2014) Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration. *PLoS Computational Biology* 10(12).

[25] Nee S (2006) Birth-Death Models in Macroevolution. *Annual Review of Ecology, Evolution, and Systematics* 37(1):1–17.

[26] Morlon H, Parsons TL, Plotkin JB (2011) Reconciling molecular phylogenies with the fossil record. *Proceedings of the National Academy of Sciences of the United States of America* 108(39):16327–16332.

[27] Stadler T (2011) Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences of the United States of America* 108(15):6187–6192.

[28] du Plessis L, Stadler T (2015) Getting to the root of epidemic spread with phylodynamic analysis of genomic data. *Trends in Microbiology* 23(7):383–386.

[29] Joy JB et al. (2016) The spread of hepatitis c virus genotype 1a in north america: a retrospective phylogenetic study. *The Lancet Infectious Diseases* 16(6):698–702.

[30] Schluter D, Pennell MW (2017) Speciation gradients and the distribution of biodiversity. *Nature* 546(7656):48–55.

[31] Louca S, Pennell MW (2020) Extant timetrees are consistent with a myriad of diversification histories. *Nature* pp. 1–4.

[32] Romero-Severson E, Skar H, Bulla I, Albert J, Leitner T (2014) Timing and order of transmission events is not directly reflected in a pathogen phylogeny. *Molecular biology and evolution* 31(9):2472–2482.

[33] Kolmogorov A (1931) Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Mathematische Annalen* 104(1):415–458.

[34] Feller W (1949) in *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability*. (University of California Press, Berkeley, California), pp. 403–432.

[35] Louca S, Pennell MW (2020) A General and Efficient Algorithm for the Likelihood of Diversification and Discrete-Trait Evolutionary Models. *Systematic Biology* 69(3):545–556.

[36] Louca S, Doebeli M (2018) Efficient comparative phylogenetics on large trees. *Bioinformatics* 34(6):1053–1055.

[37] Bouckaert R et al. (2019) BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology* 15(4):e1006650.

[38] Gavryushkina A, Welch D, Drummond AJ (2013) Recursive algorithms for phylogenetic tree counting. *Algorithms for Molecular Biology* 8(1):26.

[39] Popinga A, Vaughan T, Stadler T, Drummond AJ (2015) Inferring epidemiological dynamics with Bayesian coalescent inference: the merits of deterministic and stochastic models. *Genetics* 199(2):595–607.

[40] Louca S, Doebeli M (2018) Efficient comparative phylogenetics on large trees. *Bioinformatics* 34(6):1053–1055.

[41] Kirkpatrick M, Johnson T, Barton N (2002) General models of multilocus evolution. *Genetics* 161(4):1727–1750.

[42] Lafferty KD et al. (2015) A general consumer-resource population model. *Science* 349(6250):854–857.

[43] Kühnert D, Stadler T, Vaughan TG, Drummond AJ (2016) Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data. *Molecular Biology and Evolution* 33(8):2102–2116.

[44] Barido-Sottani J, Vaughan TG, Stadler T (2020) A Multitype Birth–Death Model for Bayesian Inference of Lineage-Specific Birth and Death Rates. *Systematic Biology*.

[45] Foote M (2003) Origination and extinction through the phanerozoic: a new approach. *The Journal of Geology* 111(2):125–148.

[46] Silvestro D, Salamin N, Antonelli A, Meyer X (2019) Improved estimation of macroevolutionary rates from fossil data using a bayesian framework. *Paleobiology* 45(4):546–570.

[47] Rabosky DL, Lovette IJ (2008) Explosive evoltuionary radiation: decreasing specaition or increasing extinction through time? *Evolution* 62(8):1866–1875.

[48] Keeling MJ, Rohani P (2008) *Modeling Infectious Diseases: In Humans and Animals*. (Princeton University Press).

[49] Rabosky DL, Lovette IJ (2008) Density-dependent diversification in North American wood warblers. *Proceedings of the Royal Society B: Biological Sciences* 275(1649):2363–2371. Publisher: Royal Society.

[50] Etienne RS, Rosindell J (2012) Prolonging the Past Counteracts the Pull of the Present: Protracted Speciation Can Explain Observed Slowdowns in Diversification. *Systematic Biology* 61(2):204–204. Publisher: Oxford Academic.

[51] Barido-Sottani J, Vaughan TG, Stadler T (2018) Detection of HIV transmission clusters from phylogenetic trees using a multi-state birth–death model. *Journal of The Royal Society Interface* 15(146):20180512.

[52] Hagen O, Hartmann K, Steel M, Stadler T (2015) Age-Dependent Speciation Can Explain the Shape of Empirical Phylogenies. *Systematic Biology* 64(3):432–440. Publisher: Oxford Academic.

[53] Leventhal GE et al. (2012) Inferring Epidemic Contact Structure from Phylogenetic Trees. *PLOS Computational Biology* 8(3):e1002413. Publisher: Public Library of Science.

[54] Vaughan TG et al. (2019) Estimating Epidemic Incidence and Prevalence from Genomic Data. *Molecular Biology and Evolution* 36(8):1804–1816.