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

Running Head: A General Birth-Death-Sampling Model

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. 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. This framework centers around a 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. We then use this framework to derive general model likelihoods for both the "single-type" case in which all lineages diversify according to the same process and the "multi-type" case, where there is variation in the process among lineages . By re-deriving existing single-type birth-death sampling models we clarify and synthesize the range of explicit and implicit assumptions made by these models.

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

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 (Drummond et al., 2003; Duffy et al., 2008). 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" (Grenfell et al., 2004; Volz, 2012). In the last two decades, there has been a tremendous amount of innovation in phylodynamic methods, and the epidemiological inferences from these models increasingly complement those from more conventional surveillance data (Volz, 2012).

Phylodynamic models can be broadly grouped into two classes. The first, based on Kingman's coalescent process (Kingman, 1982), has been historically widely used to examine changes in the historical population size of pathogens (Pybus et al., 2000; Strimmer and Pybus, 2001; Drummond et al., 2005; Volz et al., 2009). While appropriate 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 (Stadler et al., 2015; Boskova et al., 2014). The second class of models, based on the birth-death process (Kendall, 1948; Maddison et al., 2007; Stadler, 2009, 2010), makes no assumptions about sparse sampling and fully incorporates 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 (Raup, 1985; Nee et al., 1994; Morlon, 2014; Louca, 2020) and to estimate divergence times from phylogenetic data (Gernhard, 2008; Heath et al., 2014).

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 in 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 (Stadler et al., 2012, 2013) (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. A key assumption in the specification of BDS sub-models is whether all lineages are identical, diversifying according to the same process, or if rather the diversification process itself evolves (e.g., Maddison et al., 2007; FitzJohn, 2012; Stadler and Bonhoeffer, 2013; Rasmussen and Stadler, 2019; Barido-Sottani et al., 2018), with lineages belonging to one of multiple possible states each characterized by a unique process. As exemplified particularly by the range of available "single-type" models, each of these diversification processes can then be characterized by different dynamical assumptions. In the epidemiological case these assumptions specify, for example, the nature of viral transmission and the sampling procedure (Stadler et al., 2013; Kühnert et al., 2014; Gavryushkina et al., 2014). While typically not explicitly tied to mechanistic evolutionary processes, there are a similar abundance of dynamical assumptions employed in the macroevolutionary context specifying the nature of biodiversity change through time (Nee, 2006; Gernhard, 2008; Morlon et al., 2011; Stadler, 2011; Morlon, 2014; Heath et al., 2014; Louca, 2020).

This flourishing of methods and models has facilitated critical insights into epidemics (du Plessis and Stadler, 2015; Joy et al., 2016) and the origins of contemporary biodiversity (Morlon, 2014; Schluter and Pennell, 2017). However, this diversity of models has made it difficult to trace the connections between vari-

ants and to understand the precise epidemiological, evolutionary, and sampling processes that differ between them. Furthermore, despite their apparent similarities, 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 (Louca and Pennell, 2020a; Louca et al., 2021) and this is difficult to do without considering the full breadth of possible scenarios.

Here we address all of these challenges by providing a unified framework for deriving the probability of observing a given phylogeny and series of sampling events under a particular BDS model. We demonstrate the utility of five-step procedural framework to derive both a general single-type and a general multi-type model likelihood. Specifically, these general models do not assume anything about the functional forms (i.e., temporal dynamics) of the various parameters including sampling, the possibility of sampling ancestors (or not), or how the process was conditioned. Such general models may be useful for studying the mathematical properties of sets of models (Louca and Pennell, 2020a; Louca et al., 2021). However, as they are computationally cumbersome these general models will not necessarily be useful for statistical inference. Rather the procedural framework and the general model likelihood can streamline derivation of new sub-model variants adapted to specific applications.

In addition to the derivation of the general models, in the single-type case we use the procedural framework to re-derive a wide array of existing BDS models, clarifying the assumptions of each and providing likelihood expressions with standardized notation. In the multi-type case, this framework builds on recent work (Barido-Sottani et al., 2020) to provide an alternative formulation of many of the existing model likelihoods. We then use the specific case of the BiSSEness model (Magnuson-Ford and Otto, 2012; Goldberg and Igić, 2012) to illustrate how the tree likelihood can be calculated either through the traditional post-order traversal algorithm or via this five-step framework. Given our summary of single-type BDS models and the well known connections among such multi-type BDS models (for recent reviews see Morlon, 2014; Ng and Smith, 2014), our BDS framework hence provides a unification and extension to an extensive class of methods employed across epidemiological and macroevolutionary contexts.

The single-type birth-death-sampling model

Model Specification: The BDS stochastic process begins with a single lineage at time T before the present day. 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 ($0 \leq \tau \leq T$) 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 sequences (Romero-Severson et al., 2016). Throughout, we will use τ as a a general time variable and $t_{\times}$ to denote a specific time of an event $\times$ 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 ($\rho_0 > 0$). 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 Appendix. 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 (see Appendix).

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 $r(\tau)$. As with the CSAs, this arbitrary time dependence allows for the incorporation of Dirac spikes in any of these variables, for example with mass extinctions (μ) and lagerstätten in the fossil record ($\psi(1-r)$) (Magee and Höhna, 2021). 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 (such as the fossilized-birth-death process; Heath et al., 2014).

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 (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 single-type 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 $\Theta_{\text{BDS}} = \{\lambda(\tau), \mu(\tau), \psi(\tau), r(\tau), \vec{\rho}\}$.

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 panel C, 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 $g_{\text{stem}}(T)$: the probability density the stem lineage ($\text{stem} = 1$ in Figure B1 panel C) gives rise to the observed phylogeny from the origin, T, 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.

1. 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 Maddison et al. (2007). 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 (Feller, 1949; Louca and Pennell, 2020b). 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 (Louca and Pennell, 2020b).

2. 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 $g_{stem}(T)$, as follows.

3. Deriving the expression for $g_{stem}(T)$: 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.

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

4. Representing $g_{stem}(T)$ 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. 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 $\mathcal{T} = \{\vec{x}, \vec{y}, \vec{z}, T\}$. 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:

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

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:

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

Note $\Phi(0) = 1$.

5. 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 1 and listed in Table S3. The resulting likelihood of the general BDS model is:

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

Many existing models are special cases of this general BDS model

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 single-type BDS likelihood in the R package `castor` (Louca and Doebeli, 2018), including routines for maximum-likelihood fitting of 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 unified framework; for full details on each specific case, we refer readers to the Supplementary Material. While we illustrate these constraints within the single-type context, analogous assumptions can be made within the multi-type context examined in the following section. In regards to rate assumptions, many early BDS models (Stadler, 2009, 2010; Stadler et al., 2012) 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 given 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 (Stadler et al., 2012).

A similarly tractable, but more epidemiologically relevant, model is known as the "birth-death-skyline" variant (Stadler and Bonhoeffer, 2013; Gavryushkina et al., 2014), 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 Appendix). The BDS skyline model has been implemented under a variety of additional assumptions in the Bayesian phylogenetics software BEAST Bouckaert et al. (2019). The BDS skyline model has also been extended by Kuhnert et al. (Kühnert et al., 2014) 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 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, $\psi(\tau) = 0$ (Stadler, 2011).

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 (Stadler, 2009, 2011; Morlon et al., 2011) whereas past Poissonian sampling can be introduced to include the sampling of fossil data (Heath et al., 2014). In epidemiology, concerted sampling at the present day is likely biologically unrealistic (Stadler et al., 2012), though in some implementations of the models, such a sampling scheme has been imposed. These concerted sampling attempts prior to the present day as well as mass extinction events can be incorporated via the inclusion of Dirac distributions in the sampling and death rates, respectively. Finally, previous models often multiply 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 (Gavryushkina et al., 2013, 2014; Stadler, 2009), or to reflect known uncertainties. The "fossilized-birth-death" likelihood derived by Heath et al. (2014) for example, includes a factor that reflects the uncertainty in the attachment and placement of fossils on the macroevolutionary tree. This fossilized-birth-death process has been used to estimate divergence times and to model lineage diversification (Gavryushkina et al., 2017; Landis et al., 2021). Variants of the fossilized-birth-death process, for example including mass extinction events, are feasible and can be derived using our approach. We also note that models similar to the time-variable fossilized-birth-death process have been developed for cases when phylogenetic data is not available (i.e., when only including fossil occurrence data Silvestro et al., 2014; Lehtonen et al., 2017); we have not

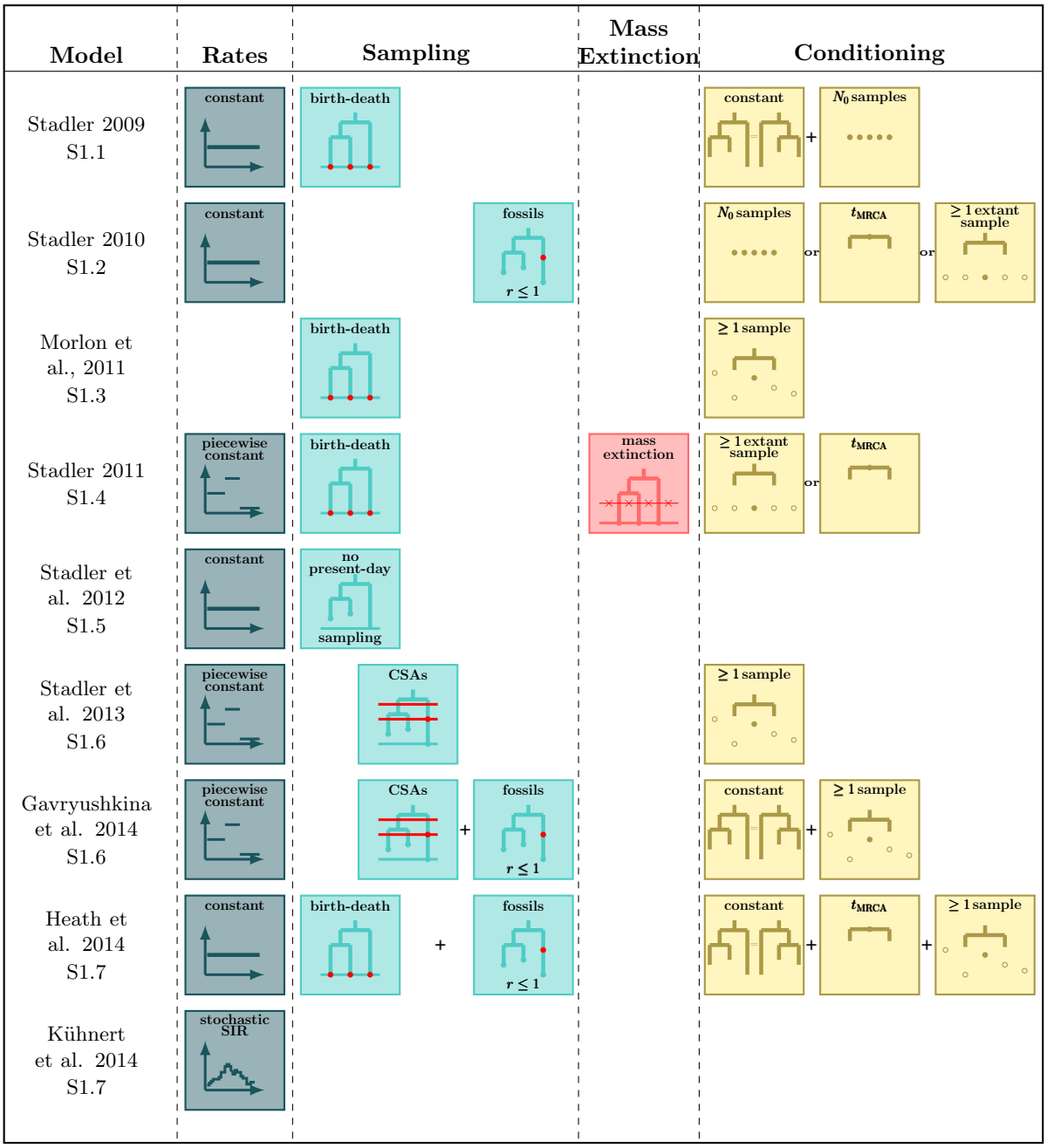

Figure 1: **Sub-model assumptions.** Rate, sampling, mass extinction, and conditioning assumptions of existing sub-models of the general time-variable BDS process. The key points are that i) each of the previously developed models we considered can be obtained by adding specific combinations of constraints 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.

investigated how these models relate to our generalized BDS model but we speculate that it would be possible to also bring these models into a common framework with those that we have discussed. 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 five-step procedure outlined above.

The multi-type birth-death-sampling model

A common extension of the single-type diversification models explored above is to consider cases where the diversification rates (λ, μ, ψ) and probabilities (r, ρ) vary among lineages as a function of a categorical "lineage type". This lineage type can be defined in terms of specific (e.g., Maddison et al., 2007; Rasmussen and Stadler, 2019) or unspecified traits (e.g., Beaulieu and O'Meara, 2016) or trait combinations (FitzJohn, 2012) (for recent reviews of these models see Morlon, 2014; Ng and Smith, 2014). Representing these lineage types as colours at nodes and along branches of the tree, we first extend the the single-type model above by deriving the likelihood of a fully coloured tree with topology $\mathcal{T}$ where the states along all edges of the phylogeny are known as given by $\mathcal{C}$. The resulting likelihood is an extension of the likelihood first developed by (Barido-Sottani et al., 2020), where the diversification rates and probabilities are allowed to vary arbitrarily through time. Following the approach of Magnuson-Ford and Otto (2012) and Goldberg and Igić (2012), the state of lineages can change either anagenetically, with a lineage of type a mutating to a type b at rate $\gamma_{a,b}(\tau)$ or cladogenetically, with a lineage of type a giving rise to a daughter lineage of type b at rate $\lambda_{a,b}(\tau)$. Lineages go extinct at a state-dependent rate $\mu_a(\tau)$ and are sampled at rate $\psi_a(\tau)$. As in the single-type model, upon sampling lineages are removed from the population with probability $r_a(\tau)$ whereas all lineages alive at the present day are sampled with a probability $\rho_a(\tau)$.

We use the five-step framework specified above for the single-type case to derive a general multi-type model likelihood of the coloured tree follows in the same manner as the single-type model (see supplementary material II). We first derive the initial value problem for the probability $g_{e,a}(\tau)$ that an edge e of type a in the tree at time τ gives rise to the subsequently observed phylogeny. The edge e here refers not to an edge in the topological tree, but to a segment of the tree all of one state between birth, sampling, or mutation events.

$$\frac{dg_{e,a}(\tau)}{d\tau} = -\left(\sum_b \lambda_{a,b} + \mu_a + \psi_a + \sum_b \gamma_{a,b}\right) g_{e,a}(\tau) + \sum_b \iota_{a,b}\lambda_{a,b} g_{e,a}(\tau) E_b(\tau)$$

$$g_{e,a}(s_e) = \begin{cases} \lambda_{a,b}(s_e) g_{e_1,a}(s_e) g_{e_2,b}(s_e) & \text{birth event } a \to a+b \\ (1 - r_a(s_e))\psi_a(s_e) g_{e_1,a}(s_e) & \text{ancestral sampling event} \\ r_a(s_e)\psi_a(s_e) + (1 - r_a(s_e))\psi_a(s_e) E_a(s_e) & \text{terminal sampling event} \\ (\gamma_{a,b}(s_e) + \lambda_{a,b} E_a)\, g_{e_1,b}(s_e) & \text{mutation/hidden birth event } a \to b \\ \rho_a & \text{sampled at present day} \end{cases} \tag{14}$$

where $\iota_{a,b}$ is an indicator variable which takes on value of 2 if $a = b$ and 1 otherwise. Importantly the differential equation for $g_{e,a}$ is linear and hence has a known general solution $g_{e,a} = g_{e,a}(s_e)\Psi(s_e, \tau)$. As in the single-type model $\Psi(s_e, \tau)$ is the probability flow (Louca and Pennell, 2020b) mapping the probability $g_{e,a}$ from the initial state at time s_e to the probability at time τ.

An analogous initial value problem can be derived for the probability $E_a(\tau)$, that a lineage of type a alive

at time τ leaves no observed descendants in the sampled tree.

$$\begin{aligned}\frac{dE_a(\tau)}{d\tau} =& -\left(\sum_b \lambda_{a,b} + \mu_a + \psi_a + \sum \gamma_{a,b}\right) E_a(\tau) + \sum_b \lambda_{a,b} E_a(\tau) E_b(\tau) + \mu_a + \sum_b \gamma_{a,b} E_b(\tau) \\ E_a(0) =& 1 - \rho_a\end{aligned} \tag{15}$$

This is a non-linear differential equation and must be solved numerically. Given the solution of $g_{e,a}$ and E_a the likelihood for the fully coloured tree is characterized by a series of critical times: first, $\vec{x}_{a,b}$ the times at which a lineage of type a gives birth to a lineage of type b, $\vec{y}_a$ the ages of tip samples of type a, $\vec{z}_a$ the ages of ancestral samples of type a, and $\vec{w}_{a,b}$ the times at which lineages are observed to transition events from type a to type b. The resulting likelihood is given by:

$$\boxed{\begin{aligned}\mathcal{L}(\mathcal{T}, \mathcal{C}|\Theta_{\text{MBDS}}) =& \mathcal{S} \times \Phi_{c^*}(T) \times \left[\prod_{a=1}^{A} \rho_a^{N_a}\right] \times \left[\prod_{a=1}^{A}\prod_{b=1}^{A}\prod_{i=1}^{I_{a,b}} \lambda_{a,b}(x_{a,b,i})\Phi_b(x_{a,b,i})\right] \\ &\times \left[\prod_{a=1}^{A}\prod_{j=1}^{J_a} \left[\psi_a(y_{a,j})(1 - r(y_{a,j}))E_a(y_{a,j}) + \psi_a(y_{a,j})r_a(y_{a,j})\right] \frac{1}{\Phi_a(y_{a,j})}\right] \\ &\times \left[\prod_{a=1}^{A}\prod_{k=1}^{K_a} \psi_a(z_{a,k})(1 - r(z_{a,k}))\right] \\ &\times \left[\prod_{a=1}^{A}\prod_{b\neq a}^{A}\prod_{l=1}^{L_{a,b}} \left[\gamma_{a,b}(w_{a,b,l}) + \lambda_{a,b}(w_{a,b,l})E_a(w_{a,b,l})\right] \frac{\Phi_a(w_{a,b,l})}{\Phi_b(w_{a,b,l})}\right]\end{aligned}} \tag{16}$$

Here $\mathcal{S}$ is an arbitrary form of conditioning as in equation (13) and $\Phi_a(\tau) = \Psi_a(\tau, 0)$, a complete list of notation is given in Table S4.

While the full colouring of the tree $\mathcal{C}$ is never known, when fitting models where the diversification parameters are a function of a measurable state of the sampled nodes, the colour at some of the tip nodes, which we denote by $\mathcal{C}_y$, and at some of the internal ancestral samples, $\mathcal{C}_z$, may be known. The likelihood of the tree topology $\mathcal{T}$ and these observed states can be calculated from equation (16) by integration.

$$\boxed{\mathcal{L}\left(\mathcal{T}, \mathcal{C}_y, \mathcal{C}_z|\Theta_{\text{MBDS}}\right) = \int_{\Omega_{\mathcal{C}|\mathcal{C}_y\mathcal{C}_z}} \mathcal{L}\left(\mathcal{T}, \mathcal{C}|\Theta_{\text{MBDS}}\right) Pr\left(\mathcal{C}|\mathcal{C}_y, \mathcal{C}_z\Theta_{\text{MBDS}}\right) d\mathcal{C}} \tag{17}$$

where $\Omega_{\mathcal{C}|\mathcal{C}_y\mathcal{C}_z}$ represents the space of all tree colourings that are consistent with the observed states. While it mathematically challenging to formulate the expression for $Pr\left(\mathcal{C}|\mathcal{C}_y, \mathcal{C}_z\Theta_{\text{MBDS}}\right)$ and hence this integral can not be done numerically it can be approximated via Monte Carlo integration. As exemplified for the BiSSEness model (Magnuson-Ford and Otto, 2012; Goldberg and Igić, 2012) in the supplementary material, one can simulate colourings within $\Omega_{\mathcal{C}|\mathcal{C}_y\mathcal{C}_z}$ using stochastic mapping (Nielsen, 2002) proportional to their respective probabilities. Averaging the likelihood given by Equation (16) across these sampled coloured trees approximates the likelihood of a tree where only the state of sampled nodes are known.

Equations (16) and (17) provide a flexible definition of the tree likelihood, not only in terms of the diversification process but with respect to the specification of lineage states. As exemplified by the work of Barido-Sottani et al. (2020), the likelihood $\mathcal{L}(\mathcal{T}, \mathcal{C}|\Theta_{MBDS})$ can be used to calculate model likelihoods where the state at the tips of the tree need not be specified a priori.

Conclusion

Here we provide a unified framework for deriving birth-death-sampling models within both single- and mutli-type contexts. We derive a general likelihood expression for each of these cases, 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 (Louca and Doebeli, 2018)), this framework can clarify the connections between various model variants, ease the development of new model variants tailored specifically to each situation, and provide a structure for understanding how results depend on model assumptions (Kirkpatrick et al., 2002; Lafferty et al., 2015; Louca and Pennell, 2020a). 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.* (Maddison et al., 2007) did for birth-death models without heterogeneous sampling. As evidence of this, 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 (Louca and Doebeli, 2018). And importantly, given the recent discovery of widespread non-identifiability in birth-death processes fit to extant-only (Louca and Pennell, 2020a) and serially-sampled (Louca et al., 2021) phylogenetic data, there is a critical need to explore a much broader range of BDS models than were previously considered and the mathematical generalization presented in our paper will be enable this.

Box 1: The connection between BDS and SIR models

The general single-type 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.

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 (Morlon et al., 2011; Rabosky and Lovette, 2008b) or discrete (Stadler, 2011; Stadler and Bonhoeffer, 2013; Gavryushkina et al., 2014; Kühnert et al., 2014) 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 (Keeling and Rohani, 2008). While some forms of diversity-dependence in diversification rates may be incorporated implicitly (Rabosky and Lovette, 2008a), explicit diversity-dependence (Etienne and Rosindell, 2012) 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 (Stadler and Bonhoeffer, 2013; Kühnert et al., 2016; Barido-Sottani et al., 2020, 2018) 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 (Hagen et al., 2015) 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 (Stadler et al., 2012).

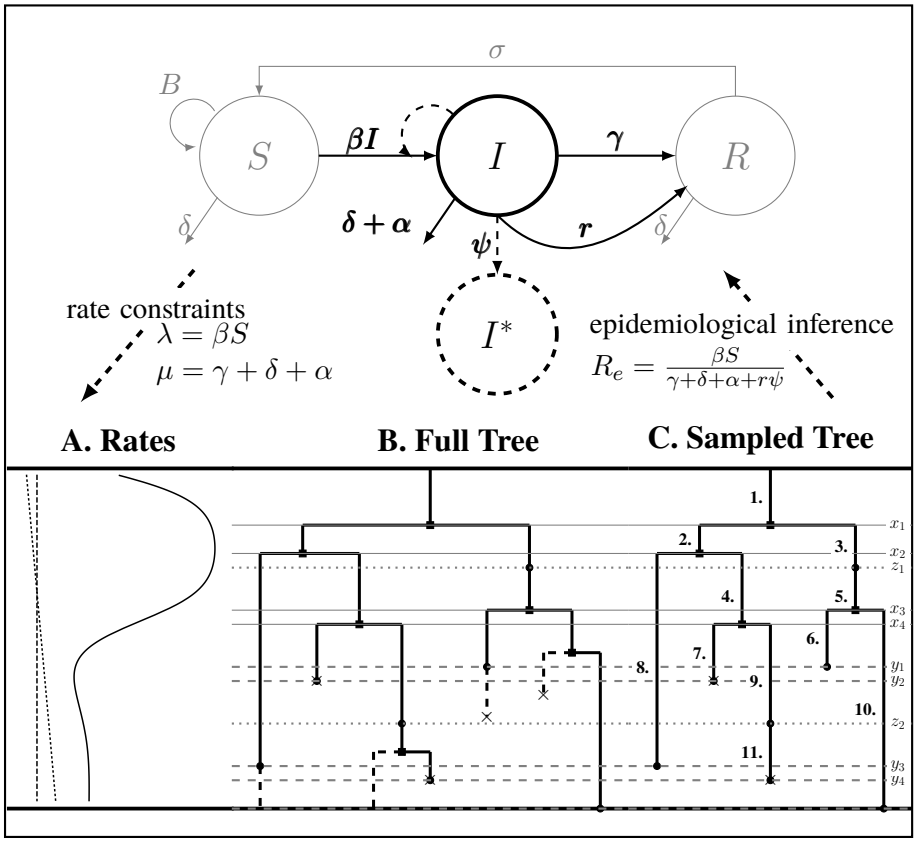

Figure B1: Top: The SIR epidemiological model. Black (gray) lines and classes represent rates and variables followed (in)directly by the BDS model. The SIR model can be used to constrain the rates of the BDS model (A.). Simulated forward in time, the result of the BDS stochastic processes is a *full 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 $\lambda(\tau) = \beta S(\tau)$ 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 (Kühnert et al., 2014) 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 (Leventhal et al., 2012) and in compartmental models in general (Vaughan et al., 2019). 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 (Stadler et al., 2012; Stadler and Bonhoeffer, 2013). Specifically, the effective reproductive rate at time τ before the present day is given by $R_e(\tau) = \frac{\lambda(\tau)}{\mu(\tau)+r(\tau)\psi(\tau)}$.

Acknowledgments

We would like to thank Sally Otto for her thoughtful comments on this work. This work was supported by a *Grant for Catalyzing Research Clusters* awarded to the UBC Biodiversity Research Centre, NSF DEB Grant #2028986 award to SL and MWP. MWP was supported by a NSERC Discovery Grant and AM was supported in part by the EEB department Postdoctoral Fellowship from the University of Toronto. AMcL was supported in part by a Canadian Institutes of Health Research (CIHR) doctoral award (#6557). JBJ is supported by a Genome Canada Bioinformatics and Computational Biology grant (287PHY), CIHR coronavirus rapid response program grant (440371) and is grateful to the British Columbia Centre for Excellence in HIV/AIDS for additional funding support.

Appendix: Adding assumptions to the general model

In this appendix, we demonstrate how one can obtain the likelihood of sub-models with different sets of assumptions by applying constraints to the general likelihood. There are four classes of assumptions that are commonly applied in epidemiological and macroevolutionary studies. First, researchers can make assumptions about the functional form of the birth, death, and sampling rates. Here we address two such unique assumptions: i) Birth, death, and sampling rates are constant (see section 1, S1.1, S1.2, S1.5); and ii) birth, death, and sampling rates are piecewise-constant functions of time (see sections 1, S1.6). The cases where birth, death, and sampling rates are defined by a stochastic or deterministic SIR model are mathematically analogous to the cases of the piecewise-constant and general time-variable models respectively. All additional constraints imposed will depend on the exact compartmental model used and hence we will not discuss them in detail in this section. The second major class of assumptions pertains to sampling. There are four such sampling assumptions: i) sampling happens only at the present day as in a birth-death model (see section 1, S1.1, S1.3, S1.4) or as implemented in the "Birth Death Skyline Contemporary" prior in the BDSKY package in BEAST; ii) the absence of concerted present-day sampling (see section 1, S1.5); iii) the inclusion of ancestral samples with sampled descendants (S1.7, and S1.6); and iv) concerted sampling attempts (CSA) during which all lineages are sampled with a given probability (see sections 1, S1.6). The third assumption class considers the presence of mass extinction events (see section 1, S1.5). The fourth and final major class of assumptions deal with the conditioning of the likelihood. The various conditioning schemes are explored in section 1 and summarized in Table S3.

Rate assumptions

Constant rates

- *Model Assumptions:*Constant diversification rates: $\lambda(t) = \lambda$, $\mu(t) = \mu$, $\psi(t) = \psi$, and constant removal probability $r(t) = r$.

- *The IVP for $g_e(\tau)$:*

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

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

- *The IVP for $E(\tau)$:*

$$\frac{dE(\tau)}{d\tau} = -(\lambda + \mu + \psi)E(\tau) + \lambda E(\tau)^2 + \mu \quad E(0) = 1 - \rho_0.$$

  In this case the IVP for $E(\tau)$ is a Bernoulli differential equation and has a known analytical solution.

As given by equation 1 in Stadler (2010) this solution is given by:

$$\begin{aligned} E(\tau) =& \frac{\lambda+\mu+\psi}{2\lambda} + \frac{c_1}{2\lambda}\frac{e^{-c_1 t}(1-c_2)-(1+c_2)}{e^{-c_1 t}(1-c_2)+(1+c_2)} \\ c_1 =& \left|\sqrt{(\lambda-\mu-\psi)^2+4\lambda\psi}\right| \quad c_2 = -\frac{\lambda-\mu-2\lambda\rho_0-\psi}{c_1}. \end{aligned} \tag{A1}$$

- *The Probability Flow:*

$$\Phi(\tau) = \exp\left[\int_0^\tau 2\lambda E(x) - (\lambda+\mu+\psi)\,dx\right].$$

- *The Likelihood:*

$$\boxed{\mathcal{L}_{\mathrm{C}} = \mathcal{S}\rho_0^{N_0}\Phi(T)\lambda^I\psi^{n+m}(1-r)^m\prod_{i=1}^{I}\Phi(x_i)\prod_{j=1}^{n}\frac{1}{\Phi(y_j)}\left[(1-r)E(y_j)+r\right]} \tag{A2}$$

Piecewise-constant rates

- Model assumptions: Divide time into $L+1$ intervals defined by *transition times* $0 = t_0 < t_1 < t_2 < \ldots < t_L < t_{L+1} = T$. Define rates and removal probabilities constant within a given interval.

$$\begin{aligned} \lambda(\tau) = \lambda_l \quad t_l < \tau \le t_{l+1} \\ \mu(\tau) = \mu_l \quad t_l < \tau \le t_{l+1} \end{aligned} \psi(\tau) = \psi_l \quad t_l < \tau \le t_{l+1} r(\tau) = r_l \quad t_l < \tau < t_{l+1}$$

- *The IVP and Solution for $g_e(\tau)$:* Given the definitions of $\lambda(\tau)$, $\mu(\tau)$, $\psi(\tau)$, and $r(\tau)$ above the IVP for $g_e(\tau)$ is identical to that given in Equations (3) and (4).

- *The IVP and Solution for $E(\tau)$:* As with $g_e(\tau)$, the IVP for $E(\tau)$ is given by Equation (8). With the piecewise-constant rate assumptions, however, the general solution for $E(\tau)$ between $t_l < \tau \le t_{l+1}$ is known (similar to Equation (A1)). Defining $E_l(\tau) = E(\tau)$ where $t_l < \tau \le t_{l+1}$ and $E_l(t_l) = E_{l-1}(t_l)$ we have:

$$\begin{aligned} E_l(\tau) =& \frac{\lambda_l+\mu_l+\psi_l}{2\lambda_l} + \frac{c_1}{2\lambda_l}\frac{e^{-c_1 t}(1-c_2)-(1+c_2)}{e^{-c_1 t}(1-c_2)+(1+c_2)} \\ c_1 =& \left|\sqrt{(\lambda_l-\mu_l-\psi_l)^2+4\lambda_l\psi_l}\right| \quad c_2 = -\frac{\lambda_l-\mu_l-2\lambda_l(1-E_l(t_l))-\psi_l}{c_1}, \end{aligned}$$

  where $E_l(t_l) = E_{l-1}(t_l)$ for $l > 0$ and $E_0(t_0) = 1-\rho_0$.

- *The Probability Flow:* We define a probability *sub-flow* within each time interval. Specifically, in the lth time interval.

$$\Phi_l(\tau) = \exp\left[\int_{t_l}^\tau 2\lambda_l E_l(x) - (\lambda_l+\mu_l+\psi_l)\,dx\right].$$

  The complete flow can be expressed as a function of the sub-flows in the following manner:

$$\begin{aligned} \Phi(\tau) =& \Phi_{L_\tau}(\tau)\prod_{l=1}^{L_\tau}\Phi_{l-1}(t_l) \\ \Phi(t_k) =& \underbrace{\Phi_k(t_k)}_{1}\prod_{l=1}^{k}\Phi_{l-1}(t_l) = \prod_{l=1}^{k}\Phi_{l-1}(t_l), \end{aligned} \tag{A3}$$

where L_t is the index of the time t_l at or after time t, i.e. the largest index such that $t_l \leq \tau$.

- *The Likelihood:* Given these piecewise definitions we substitute them into the general BDS likelihood (13).

$$\mathcal{L}_{PC} = \mathcal{S} \underbrace{\rho_0^{N_0}}_{\substack{\text{extant}\\\text{tips}}} \underbrace{\Phi_L(T) \prod_{l=1}^{L} \Phi_{l-1}(t_l)}_{\text{root}} \times \underbrace{\prod_{i=1}^{I} \left[\lambda_{L_{x_i}} \Phi_{L_{x_i}}(x_i) \prod_{l=1}^{L_{x_i}} \Phi_{l-1}(t_l) \right]}_{\text{births}}$$
$$\times \underbrace{\prod_{j=1}^{n} \frac{\psi_{L_{y_j}} \left((1 - r_{L_{y_j}}) E_{L_{y_j}}(y_j) + r_{L_{y_j}} \right)}{\Phi_{L_{y_j}} \prod_{l=1}^{L_{y_j}} \Phi_{l-1}(t_l)}}_{\substack{\text{extinct}\\\text{tips}}} \times \underbrace{\prod_{k=1}^{m} \psi_{L_{z_k}} \left(1 - r_{L_{z_k}} \right)}_{\substack{\text{ancestral}\\\text{samples}}},$$

where we use PC to denote the piecewise-constant assumption.

We can simplify several of these products. Let α_l be the number of birth events $\geq t_l$ and σ_l the number of sampling events $\geq t_l$.

$$\begin{aligned} \prod_{i=1}^{I} \prod_{l=1}^{L_{x_i}} \Phi_{l-1}(t_l) &= \prod_{l=1}^{L} \left[\Phi_{l-1}(t_l) \right]^{\alpha_l} \\ \prod_{j=1}^{n} \prod_{l=1}^{L_{y_j}} \frac{1}{\Phi_{l-1}(t_l)} &= \prod_{l=1}^{L} \left[\Phi_{l-1}(t_l) \right]^{-\sigma_l}. \end{aligned} \tag{A4}$$

Let n_l be the number of observed lineages alive at time t_l. Because the number of observed lineages increases with each birth and decreases with each sampled tip, counting the root we have $n_l = \alpha_l - \sigma_l + 1$. Substituting the expressions for the into the likelihood and using the definition of n_l we have:

$$\boxed{\begin{aligned} \mathcal{L}_{PC} =& \mathcal{S} \rho_0^{N_0} \Phi_L(T) \prod_{i=1}^{I} \lambda_{L_{x_i}} \Phi_{L_{x_i}}(x_i) \\ & \times \prod_{j=1}^{n} \frac{\psi_{L_{y_j}}}{\Phi_{L_{y_j}}} \left((1 - r_{L_{y_j}}) E_{L_{y_j}}(y_j) + r_{L_{y_j}} \right) \prod_{k=1}^{m} \psi_{L_{z_k}} \left(1 - r_{L_{z_k}} \right) \\ & \times \prod_{l=1}^{L} \Phi_{l-1}(t_l)^{n_l}. \end{aligned}} \tag{A5}$$

Sampling assumptions

Birth-death models

- *Model Assumptions:* The birth-death model assumes that $\psi(\tau) = 0$. Note that the probability of sampling a lineage given it is alive at the present day remains as ρ_0 (incomplete sampling).

- *IVP for $g_e(\tau)$:*

$$\begin{aligned} \frac{dg_e(\tau)}{d\tau} &= -\left(\lambda(\tau) + \mu(\tau)\right) g_e(\tau) + 2\lambda(\tau) g_e(\tau) E(\tau) \\ g_e(s_e) &= \begin{cases} \lambda(s_e) g_{e1}(s_e) g_{e2}(s_e) & \text{birth event giving rise to edges e1 and e2} \\ \rho_0 & s_e = 0\text{, edge sampled at present day} \end{cases} \end{aligned}$$

- *IVP for $E(\tau)$:*

$$\frac{dE(\tau)}{d\tau} = -(\lambda(\tau) + \mu(\tau))E(\tau) + \lambda(\tau)E(\tau)^2 + \mu(\tau) \quad E(0) = 1 - \rho_0.$$

Note in this case $E(\tau)$ equals $\hat{E}(\tau)$, the probability a lineage leaves no sampled extant descendants. As demonstrated by Morlon et al. (2011) there exists a general solution to this initial value problem, see section S1.3 for more details. This general solution is given by:

$$E(\tau) = 1 - \frac{\rho_0 \exp\left[\int_0^\tau (\lambda(u) - \mu(u))\, du\right]}{1 + \int_0^\tau \rho_0 \exp\left[\int_0^x (\lambda(u) - \mu(u))\, du\right] dx}.$$

- *The Probability Flow:* From (Morlon et al., 2011), the probability flow can be written as the following:

$$\Phi(\tau) = \exp\left[\int_0^\tau (\lambda(\sigma) - \mu(\sigma))\, d\sigma\right] \left[1 + \frac{\int_0^\tau \rho_0 \lambda(u) \exp\left[\int_0^u (\lambda(\sigma) - \mu(\sigma))\, d\sigma\right] du}{1 + \rho_0}\right]^{-2}.$$

- *The Likelihood:*

$$\boxed{\mathcal{L}_{BD} = \mathcal{S}\rho_0^{N_0}\Phi(T)\prod_{i=1}^{I}\lambda(x_i)\Phi(x_i)} \tag{A6}$$

No sampling at the present

Here we consider the case when $\rho_0 = 0$. The likelihood follows exactly as in the general model case. The resulting likelihood expression is given by:

$$\boxed{\begin{aligned}\mathcal{L}_{\rho_0=0} =& \mathcal{S}\Phi(T)\prod_{i=1}^{I}\lambda(x_i)\Phi(x_i) \\ &\times \prod_{j=1}^{n}\frac{\psi(y_j)}{\Phi(y_j)}\left[(1 - r(y_j))E(y_j) + r(y_j)\right]\prod_{k=1}^{m}\psi(z_k)\left(1 - r(z_k)\right).\end{aligned}} \tag{A7}$$

Note that in this case $I = n - 1$.

Concerted sampling attempts

- *Model Assumptions:*Here we introduce L concerted sampling attempts (CSA) at known points in time, $t_l \quad l \in \{1, 2, ...L\}$. Like the CSA at the present day, and in contrast to the background Poissonian sampling rate, during the CSA at time t_l every lineages is sampled with a fixed probability ρ_l. In the derivation of the likelihood below, we must distinguish between three different sampling event types. First, *past Poissonian sampling events* are those that do not occur during CSAs. Second, *past concerted sampling events* are those that occur during a CSA at time $t_l \quad l \in \{1, 2, ..., L\}$. Finally, *present concerted sampling events* are those that occur at the present day $\tau = 0$. Past concerted sampling attempts can be included in the general model above by adding L Dirac distributions to the Poisson sampling rate function. Namely,

$$\psi(\tau) = \bar{\psi}(\tau) + \sum_{l=1}^{L}\left[w_l * \delta(\tau - t_l)\right], \tag{A8}$$

where $\bar{\psi}(\tau)$ is the background Poissonian sampling rate and $w_l = -ln\,(1-\rho_l)$. The definition of w_l comes from solving the CDF of the exponentially distribution for the 'effective sampling rate' such that the probability of a lineage being sampled is ρ_l.

- *IVP for $g_e(\tau)$:*

$$\frac{dg_e(\tau)}{d\tau} = -\left(\lambda(\tau)+\mu(\tau)+\psi(\tau)\right)g_e(\tau) + 2\lambda(\tau)g_e(\tau)E(\tau)$$

$$g_e(s_e) = \begin{cases} \lambda(s_e)g_{e1}(s_e)g_{e2}(s_e) & \text{birth event giving rise to edges e1 and e2} \\ (1-r(s_e))\bar{\psi}(s_e)g_{e1}(s_e) & \text{Poisson ancestral sampling event} \\ \bar{\psi}(s_e)r(s_e) + \bar{\psi}(s_e)(1-r(s_e))E(s_e) & \text{Poisson terminal sampling event} \\ (1-r(t_l))\rho_l g_{e1}(t_l) & \text{ancestral sample at } t_l \\ \rho_l r(t_l) + \rho_l(1-r(t_l))E(t_l) & \text{terminal sample at } t_l \\ \rho_0 & s_e = 0\text{, edge sampled at present day} \end{cases}$$

The solution to $g_e(\tau)$ is given by Equations (5) and (6).

- *IVP for $E(\tau)$:* As with $g_e(\tau)$, the IVP for $E(\tau)$ is identical to that given for the general model in Equation (8). Except in rare cases the IVP must be solved numerically hence requiring numerical integration over Dirac distributions which can prove to be problematic.

  Note however, that when examining the integrals over the CSAs, a priori, it is a matter of convention whether the Dirac distribution should be considered as "integrated over" when located at the upper integration bound $\int_a^b \delta(s-b)ds = 1$ or at the lower integration bound $\int_a^b \delta(s-a)ds = 1$. Whichever convention we chose, we must rigorously obey it so that the ratio $\Phi(t)/\Phi(s)$ correctly evaluates to $\Psi(s,t)$ whenever $s \leq t$. Using the former convention, we can rewrite the probability $E(t_l)$ at each concerted sampling time t_l as:

$$E(t_l) = E(t_l^-)e^{w_l} = E(t_l^-)(1-\rho_l),$$

  where t_l^- denotes the limit as time approaches t_l from below. Hence the probability $E(\tau)$ at any time τ can be evaluated numerically by considering the dynamics between successive CSAs and at each CSA separately.

- *The Probability Flow:* The probability flow is given by:

$$\Phi(\tau) = \exp\left[\int_0^\tau 2\lambda(x)E(x) - \left(\lambda(x)+\mu(x)+\bar{\psi}(x)+\sum_{l=1}^{L} w_l\delta(x-t_l)\right)dx\right].$$

  As with $E(\tau)$ integration over the dirac distributions can be problematic and hence we rewrite this expression separating out these terms. Let L_τ be the oldest CSA occurring at or after time τ, i.e. the largest index for which $t_l \leq \tau$.

$$\begin{aligned}\Phi(\tau) &= \exp\left[\int_0^\tau 2\lambda(x)E(x) - \left(\lambda(x)+\mu(x)+\bar{\psi}(x)+\sum_{l=1}^{L_\tau} w_l\delta(x-t_l)\right)dx\right] \\ &= \exp\left[\int_0^\tau 2\lambda(x)E(x) - \left(\lambda(x)+\mu(x)+\bar{\psi}(x)\right)dx\right]\prod_{l=1}^{L_\tau} e^{-w_l} \qquad \text{(A9)} \\ &= \exp\left[\int_0^\tau 2\lambda(x)E(x) - \left(\lambda(x)+\mu(x)+\bar{\psi}(x)\right)dx\right]\prod_{l=1}^{L_\tau}(1-\rho_l).\end{aligned}$$

We define:

$$\bar{\Phi}(\tau) = \exp\left[\int_0^{\tau} 2\lambda(x)E(x) - \Big(\lambda(x) + \mu(x) + \bar{\psi}(x)\Big)\, dx\right], \tag{A10}$$

which means that we can rewrite Equation (A9) as:

$$\Phi(\tau) = \bar{\Phi}(\tau) \prod_{l=1}^{L_\tau} (1 - \rho_l). \tag{A11}$$

- *The Likelihood:* The edge representation of g_{stem} is given by:

$$\begin{aligned} g_{stem}(T) =& \rho_0^{N_0} \prod_{i=1}^{L} \lambda(x_i) \prod_{j=1}^{n} \psi(y_j) \left[(1 - r(y_j))E(y_j) + r(y_j)\right] \prod_{k=1}^{m} \psi(z_j)(1 - r(y_j)) \\ &\times \prod_{l=1}^{L} \rho_l \left[(1 - r_l)E(t_l) + r_l\right]^{N_l} \prod_{l=1}^{L} \left[\rho_l(1 - r_l)\right]^{M_l} \prod_{edges} \Psi(s_e, t_e). \end{aligned}$$

The critical time representation of g_{stem} is given by:

$$\begin{aligned} g_{stem}(T) = \; & \underbrace{\rho_0^{N_0}}_{\text{extant tips}} \underbrace{\bar{\Phi}(T) \prod_{l=1}^{L} (1 - \rho_l)}_{\text{root}} \underbrace{\prod_{i=1}^{I} \lambda(x_i) \bar{\Phi}(x_i) \left[\prod_{l=1}^{L_{x_i}} (1 - \rho_l)\right]}_{\text{births}} \\ & \times \underbrace{\prod_{j=1}^{n} \frac{\psi(y_j)}{\bar{\Phi}(y_j)} \left[(1 - r(y_j))E(y_j) + r(y_j)\right] \left[\prod_{l=1}^{L_{y_j}} \frac{1}{1 - \rho_l}\right]}_{\text{Pois. extinct tips}} \underbrace{\prod_{k=1}^{m} \psi(z_k)(1 - r(z_k))}_{\text{Pois. ances. samples}} \\ & \times \underbrace{\prod_{l=1}^{L} \left(\frac{\rho_l}{\bar{\Phi}(t_l)} \left[(1 - r_l)E(t_l) + r_l\right]\right)^{N_l} \left[\prod_{j=1}^{l} \frac{1}{(1 - \rho_j)^{N_l}}\right]}_{\text{CSA extinct tips}} \underbrace{\prod_{l=1}^{L} \left[\rho_l(1 - r_l)\right]^{M_l}}_{\text{CSA ances. samples}}, \end{aligned}$$

where N_l is the number of tip samples (samples without descendants) obtained during the lth CSA and M_l is the number of ancestral samples (sequences with descendants). By changing how we enumerate birth, death, and sampling events we can greatly simplify this likelihood. First, let α_l be the number of branching events at or before the the lth CSA. In other words, α_l is the number of branching events if the tree were trimmed at the lth CSA. Then:

$$\prod_i \left[\prod_{l=1}^{L_{x_i}} (1 - \rho_l)\right] = \prod_{l=1}^{L} (1 - \rho_l)^{\alpha_l}. \tag{A12}$$

Second, let σ_l be the number of past Poissonian sampling events before time t_l. Then:

$$\prod_j^n \left[\prod_{l=1}^{L_{y_j}} \frac{1}{(1 - \rho_l)}\right] = \prod_{l=1}^{L} \frac{1}{(1 - \rho_l)^{\sigma_l}}. \tag{A13}$$

Finally, let β_l be the number of past lineages sampled during a CSA at or before the CSA at time t_l. Hence, $\beta_l = N_l + N_{l+1} + ... + N_L$. Then:

$$\prod_{l=1}^{L} \left[\prod_{j=1}^{l} \frac{1}{(1 - \rho_j)^{N_l}}\right] = \prod_{l=1}^{L} \frac{1}{(1 - \rho_l)^{\beta_l}}. \tag{A14}$$

The likelihood hence simplifies to:

$$
\begin{aligned}
g_{stem}(T) =& \rho_0^{N_0} \bar{\Phi}(T) \prod_{l=1}^{L} (1-\rho_l)^{\alpha_l - \beta_l - \sigma_l + 1} \prod_{i=1}^{I} \lambda(x_i) \bar{\Phi}(x_i) \\
& \times \prod_{j=1}^{n} \frac{\psi(y_j)}{\bar{\Phi}(y_j)} \left[(1-r(y_j))E(y_j) + r(y_j) \right] \prod_{k=1}^{m} \psi(z_k)(1-r(z_k)) \\
& \times \prod_{l=1}^{L} \left(\frac{\rho_l}{\bar{\Phi}(t_l)} \left[(1-r_l)E(t_l) + r_l \right] \right)^{N_l} \prod_{l=1}^{L} \left[\rho_l (1-r_l) \right]^{M_l}.
\end{aligned}
$$

Let n_l be the number of lineages that cross t_l, i.e., the number of lineages alive at time t_l with sampled descendants at some younger age. Note that by this definition $n_0 = 0$. Then $b_l + \sigma_l + n_l$ is the number of tips in the tree had it been trimmed at age t_l whereas α_l is the number of branching events. Therefore we must have $\alpha_l = b_l + \sigma_l + n_l - 1$. This allows us to simplify the conditioned likelihood given below.

$$
\boxed{
\begin{aligned}
\mathcal{L}_{\text{CSA}} =& \mathcal{S} \rho_0^{N_0} \bar{\Phi}(T) \prod_{l=1}^{L} (1-\rho_l)^{n_l} \prod_{i=1}^{I} \lambda(x_i) \bar{\Phi}(x_i) \\
& \times \prod_{j=1}^{n} \frac{\psi(y_j)}{\bar{\Phi}(y_j)} \left[(1-r(y_j))E(y_j) + r(y_j) \right] \prod_{k=1}^{m} \psi(z_k)(1-r(z_k)) \\
& \times \prod_{l=1}^{L} \left(\frac{\rho_l}{\bar{\Phi}(t_l)} \left[(1-r_l)E(t_l) + r_l \right] \right)^{N_l} \prod_{l=1}^{L} \left[\rho_l (1-r_l) \right]^{M_l}
\end{aligned}
} \tag{A15}
$$

Mass extinction

- *Model Assumptions:* In addition to the Poisson birth death and sampling events considered in the general model, there are L mass extinctions events occurring at times $t_1 > t_2 > ..t_L$. During the lth mass extinction event each lineage goes extinct with probability ν_l. As with concerted sampling such mass extinction events can be introduced into the model by adding a set of dirac-delta functions to the Poisson death rate, $\bar{\mu}(\tau)$.

$$
\mu(\tau) = \bar{\mu}(\tau) + \sum_{l=1}^{L} m_l \delta(\tau - t_l), \tag{A16}
$$

  where $m_l = -ln(1-\nu_l)$.

- *IVP for $g_e(\tau)$:* The initial value problem for $g_e(\tau)$ is identical to that given in equation by Equations (3) and (4) except that μ is now includes the mass extinction events.

- *IVP for $E(\tau)$:* The IVP for $E(\tau)$ to that given by Equation (8) except where the extinction rate is given by Equation (A16). The solution to $E(\tau)$ is obtained by numerical integration. Given the dirac-delta functions this numerical integration can be carried out in a piecewise manner integrating separately between and over each mass extinction event. Defining $E(t_l^-)$ as the solution up to but not including the mass extinction event at time t_l we have:

$$
E(t_l) = (1-\nu_l)E(t_l^-) + \nu_l.
$$

The first term reflects the probability that a lineage that does not go extinct during the lth mass extinction event leaves no observable offspring (with probability $E(t_l^-)$) whereas the second term reflects the fact that all lineages that go extinct during the lth mass extinction leave no observed descendants with probability 1.

- *The Probability Flow:* The solution to the IVP is once again given by $g_e(\tau) = g_e(s_e)\Psi(s_e, \tau) = g_e(s_e)\frac{\Phi(\tau)}{s_e}$ where:

$$\Phi(\tau) = \exp\left[\int_0^\tau 2\lambda(x)E(x) - \left(\lambda(x) + \bar{\mu}(x) + \sum_{l=1}^{L} m_l\delta(x - t_l) + \psi(x)\right) dx\right].$$

As with the CSAs, let L_τ be the last index l such that $t_l < \tau$. We can separate out the mass extinction terms in the following way.

$$\begin{aligned}\Phi(\tau) =& \exp\left[\int_0^\tau 2\lambda(x)E(x) - (\lambda(x) + \bar{\mu}(x) + \psi(x))\, dx\right] \prod_{l=1}^{L_\tau} e^{-m_l} \\ =& \exp\left[\int_0^\tau 2\lambda(x)E(x) - (\lambda(x) + \bar{\mu}(x) + \psi(x))\, dx\right] \prod_{l=1}^{L_\tau} (1 - \nu_l) \\ =& \bar{\Phi}(\tau) \prod_{l=1}^{L_\tau} (1 - \nu_l)\,,\end{aligned}$$

where $\bar{\Phi}(\tau)$ is defined as in Equation (A10).

- *The Likelihood:* Given these initial value problems the likelihood follows as in the general model.

$$\begin{aligned}\mathcal{L}_{ME} =& \mathcal{S}\rho_0^{N_0}\bar{\Phi}(T) \prod_{l=1}^{L}(1 - \nu_l) \prod_{i=1}^{I} \left[\lambda(x_i)\bar{\Phi}(x_i) \prod_{l=1}^{L_{x_i}}(1 - \nu_l)\right] \\ &\times \prod_{j=1}^{n} \left[\frac{\psi(y_j)\left[(1 - r(y_j))E(y_j) + r(y_j)\right]}{\bar{\Phi}(y_j)\prod_{l=1}^{L_{y_j}}(1 - \nu_l)}\right] \prod_{k=1}^{m} \psi(z_k)(1 - r(z_k)).\end{aligned}$$

As with the CSAs we can use relations analogous to Equations A12-A14 to rewrite the likelihood:

$$\boxed{\begin{aligned}\mathcal{L}_{ME} =& \mathcal{S}\rho_0^{N_0}\bar{\Phi}(T) \prod_{l=1}^{L}(1 - \nu_l)^{n_l} \\ &\times \prod_{i=1}^{I} \lambda(x_i)\bar{\Phi}(x_i) \prod_{j=1}^{n} \frac{\psi(y_j)}{\bar{\Phi}(y_j)} \left[(1 - r(y_j))E(y_j) + r(y_j)\right] \prod_{k=1}^{m} \psi(z_k)(1 - r(z_k)),\end{aligned}} \tag{A17}$$

where n_l is defined as before as the number of lineages present at time t_l.

Alternative conditioning

Table S3 lists a number of possible conditionings, $\mathcal{S}$ that can be applied to the tree likelihood. First, is the trivial case of no conditioning $\mathcal{S}_0 = 1$ which gives the likelihood of the observed tree including the stem edge

between time T and t_{MRCA}. The likelihood of the phylogeny excluding the stem edge, i.e., conditioning of the t_{MRCA}, can be obtained by setting $\mathcal{S} = \mathcal{S}_1 = \frac{\Phi(x_1)}{\Phi(T)}$. Recall that the elements of $\vec{x}$ are ordered such that $x_1 = t_{\text{MRCA}}$ is the first (oldest) birth event.

Acknowledging that one would not reconstruct a phylogeny without any sampled lineages, we can condition the likelihood on observing at least one sampled lineage (either at or before the present day) given the time of origin, $\mathcal{S}_2 = \frac{1}{1-E(T)}$. Or as with $\mathcal{S}_1$, conditioning on at least one sampled lineage given the t_{MRCA}. In order to have at least one sampled lineage *and* a most recent common ancestor, however, each daughter lineage of the common ancestor must have at least one descendent. Hence we have $\mathcal{S}_3 = \frac{\Phi(x_1)}{\Phi(T)} \frac{1}{(1-E(x_1))^2}$. The general birth-death-sampling model assumes that all lineages alive at the present day are sampled with probability ρ_0. As with concerted sampling attempts (CSAs) prior to the present day, this present day CSA may include the sampling of multiple lineages as well as possibly resulting in no sampled lineages. As with $\mathcal{S}_2$ and $\mathcal{S}_3$ we can condition the tree likelihood on observing at least one extant lineage at the present day. To do so, we define $\hat{E}(\tau) = E(\tau|\psi = 0)$, the probability that a lineage alive at time τ has no extant descendants. Conditioning on the time of origin we have: $\mathcal{S}_4 = \frac{1}{1-\hat{E}(T)}$. Conditioning on the time of the most recent common ancestor we have: $\mathcal{S}_5 = \frac{\Phi(x_1)}{\Phi(T)} \frac{1}{2\left(1-\hat{E}(x_1)\right)(1-E(x_1))}$, where now at least one of the two daughter lineages of the common ancestor has a present day sample. In many cases $\mathcal{S}_5$ is modified, however, to condition on *both* daughter lineages having an extant sampled descendent: $\mathcal{S}_5' = \frac{\Phi(x_1)}{\Phi(T)} \frac{1}{\left(1-\hat{E}(x_1)\right)^2}$.

As an alternative to conditioning on at least one extant sampled descent, tree likelihoods can be conditioned on having *exactly* N_0 sampled (extant) descendants. Let $\hat{E}_{N_0}(\tau)$ be the probability a lineage alive at time τ has exactly N_0 descendants. Although a general expression for $\hat{E}_{N_0}(\tau)$ is unknown, in the case of constant birth, death, and sampling rates (the case in which this form of conditioning has been applied), the expression for $\hat{E}_{N_0}(\tau)$ is given by (Gernhard, 2008; Kendall, 1948) and **Theorem 3.3** by Stadler Stadler (2010):

$$\begin{aligned}
\hat{E}_{N_0}(\tau) =& \rho_0 \hat{\Phi}(\tau) \left(\frac{\lambda}{\mu} \hat{E}(\tau)\right)^{N_0-1} \\
\hat{E}_{N_0}(\tau) =& \rho_0 \hat{\Phi}(\tau) \left(\frac{\rho_0 \lambda \left(1 - e^{-(\lambda-\mu)t}\right)}{\lambda\rho_0 + (\lambda(1-\rho_0) - \mu)\, e^{-(\lambda-\mu)t}}\right)^{N_0-1},
\end{aligned}$$

where, like $\hat{E}$, $\hat{\Phi}$ is given by Equation (10) evaluated with where $\psi = 0$. Given the time of origin we can condition on observing exactly N_0 lineages by setting $\mathcal{S} = \mathcal{S}_6 = \frac{1}{\hat{E}_{N_0}(T)}$. When t_{MRCA} is given instead, then the number of descendants of the two daughter lineages must add up to N_0 while both daughter lineages must still have at least one descendant(see Stadler (2010) **Corollary 3.9**).

$$\begin{aligned}
\mathcal{S} = \mathcal{S}_7 =& \frac{\Phi(x_1)}{\Phi(t_{\text{or}})} \left(\sum_{i=1}^{N_0-1} \hat{E}_i(x_1) \hat{E}_{N_0-i}(x_1)\right)^{-1} \\
=& \frac{\Phi(x_1)}{\Phi(t_{\text{or}})} \left[(N_0-1)(\rho_0 \hat{\Phi}(x_1))^2 \left(\frac{\rho_0 \lambda \left(1 - e^{-(\lambda-\mu)t}\right)}{\lambda\rho_0 + (\lambda(1-\rho_0) - \mu)\, e^{-(\lambda-\mu)t}}\right)^{N_0-2}\right]^{-1}.
\end{aligned}$$

While early BDS models often employed such conditioning (Stadler, 2009, 2010), this form of conditioning has not been employed in many later models perhaps because the biological justification for such conditioning is vague.

The final form of conditioning used in the literature, which we will represent simply as $\mathcal{S}_8$, is the multiplication of the BDS likelihood by a constant to account for the enumeration over the possible indistinguishable

representations of a given tree. The value of this constant depends on whether the tree considered is "labeled" and "oriented" (Gavryushkina et al., 2013) and whether, as in the derivation here, the vector of birth events, $\vec{x}$, is (un)ordered. Inclusion of such a constant should have no effect on the maximum likelihood inference of the model parameters given a specified tree. In cases where the constant is a function of the critical times (Heath et al., 2014), it can influence the inference when the parameters and the tree are jointly estimated.

References

Barido-Sottani, J., T. G. Vaughan, and T. Stadler. 2018. Detection of HIV transmission clusters from phylogenetic trees using a multi-state birth–death model. Journal of The Royal Society Interface 15:20180512.

Barido-Sottani, J., T. G. Vaughan, and T. Stadler. 2020. A Multitype Birth–Death Model for Bayesian Inference of Lineage-Specific Birth and Death Rates. Systematic Biology .

Beaulieu, J. M. and B. C. O'Meara. 2016. Detecting Hidden Diversification Shifts in Models of Trait-Dependent Speciation and Extinction. Systematic Biology 65:583–601.

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

Bouckaert, R., T. G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, A. Gavryushkina, J. Heled, G. Jones, D. Kühnert, N. De Maio, M. Matschiner, F. K. Mendes, N. F. Müller, H. A. Ogilvie, L. du Plessis, A. Popinga, A. Rambaut, D. Rasmussen, I. Siveroni, M. A. Suchard, C.-H. Wu, D. Xie, C. Zhang, T. Stadler, and A. J. Drummond. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Computational Biology 15:e1006650.

Drummond, A. J., O. G. Pybus, A. Rambaut, R. Forsberg, and A. G. Rodrigo. 2003. Measurably evolving populations. Trends in Ecology & Evolution 18:481–488.

Drummond, A. J., A. Rambaut, B. Shapiro, and O. G. Pybus. 2005. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. Molecular Biology and Evolution 22:1185–1192.

du Plessis, L. and T. Stadler. 2015. Getting to the root of epidemic spread with phylodynamic analysis of genomic data. Trends in Microbiology 23:383–386.

Duffy, S., L. A. Shackelton, and E. C. Holmes. 2008. Rates of evolutionary change in viruses: Patterns and determinants. Nature Reviews Genetics 9:267–276.

Etienne, R. S. and J. Rosindell. 2012. Prolonging the Past Counteracts the Pull of the Present: Protracted Speciation Can Explain Observed Slowdowns in Diversification. Systematic Biology 61:204–204.

Feller, W. 1949. Proceedings of the [First] Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley California.

FitzJohn, R. G. 2012. Diversitree : Comparative phylogenetic analyses of diversification in R: *Diversitree*. Methods in Ecology and Evolution 3:1084–1092.

Gavryushkina, A., T. A. Heath, D. T. Ksepka, T. Stadler, D. Welch, and A. J. Drummond. 2017. Bayesian Total-Evidence Dating Reveals the Recent Crown Radiation of Penguins. Systematic Biology 66:57–73.

Gavryushkina, A., D. Welch, and A. J. Drummond. 2013. Recursive algorithms for phylogenetic tree counting. Algorithms for Molecular Biology 8:26.

Gavryushkina, A., D. Welch, T. Stadler, and A. J. Drummond. 2014. Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration. PLoS Computational Biology 10.

Gernhard, T. 2008. The conditioned reconstructed process. Journal of Theoretical Biology 253:769–778.

Goldberg, E. E. and B. Igić. 2012. Tempo and Mode in Plant Breeding System Evolution. Evolution 66:3701–3709.

Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science (New York, N.Y.) 303:327–332.

Hagen, O., K. Hartmann, M. Steel, and T. Stadler. 2015. Age-Dependent Speciation Can Explain the Shape of Empirical Phylogenies. Systematic Biology 64:432–440.

Heath, T. A., J. P. Huelsenbeck, and T. Stadler. 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences 111:E2957–E2966.

Joy, J. B., R. M. McCloskey, T. Nguyen, R. H. Liang, Y. Khudyakov, A. Olmstead, M. Krajden, J. W. Ward, P. R. Harrigan, J. S. G. Montaner, and A. F. Y. Poon. 2016. The spread of hepatitis C virus genotype 1a in North America: A retrospective phylogenetic study. The Lancet. Infectious Diseases 16:698–702.

Keeling, M. J. and P. Rohani. 2008. Modeling Infectious Diseases: In Humans and Animals. Princeton University Press.

Kendall, D. G. 1948. On the Generalized "Birth-and-Death" Process. Annals of Mathematical Statistics 19:1–15.

Kingman, J. F. C. 1982. On the Genealogy of Large Populations. Journal of Applied Probability 19:27–43.

Kirkpatrick, M., T. Johnson, and N. Barton. 2002. General models of multilocus evolution. Genetics 161:1727–1750.

Kühnert, D., T. Stadler, T. G. Vaughan, and A. J. Drummond. 2014. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. Journal of The Royal Society Interface 11:20131106.

Kühnert, D., T. Stadler, T. G. Vaughan, and A. J. Drummond. 2016. Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data. Molecular Biology and Evolution 33:2102–2116.

Lafferty, K. D., G. DeLeo, C. J. Briggs, A. P. Dobson, T. Gross, and A. M. Kuris. 2015. A general consumer-resource population model. Science 349:854–857.

Landis, M. J., D. A. R. Eaton, W. L. Clement, B. Park, E. L. Spriggs, P. W. Sweeney, E. J. Edwards, and M. J. Donoghue. 2021. Joint Phylogenetic Estimation of Geographic Movements and Biome Shifts during the Global Diversification of Viburnum. Systematic Biology 70:67–85.

Lehtonen, S., D. Silvestro, D. N. Karger, C. Scotese, H. Tuomisto, M. Kessler, C. Peña, N. Wahlberg, and A. Antonelli. 2017. Environmentally driven extinction and opportunistic origination explain fern diversification patterns. Scientific Reports 7:4831.

Leventhal, G. E., R. Kouyos, T. Stadler, V. von Wyl, S. Yerly, J. Böni, C. Cellerai, T. Klimkait, H. F. Günthard, and S. Bonhoeffer. 2012. Inferring Epidemic Contact Structure from Phylogenetic Trees. PLOS Computational Biology 8:e1002413.

Louca, S. 2020. Simulating trees with millions of species. Bioinformatics .

Louca, S. and M. Doebeli. 2018. Efficient comparative phylogenetics on large trees. Bioinformatics 34:1053–1055.

Louca, S., A. McLaughlin, A. MacPherson, J. B. Joy, and M. W. Pennell. 2021. Fundamental identifiability limits in molecular epidemiology. bioRxiv Page 2021.01.18.427170.

Louca, S. and M. W. Pennell. 2020a. Extant timetrees are consistent with a myriad of diversification histories. Nature Pages 1–4.

Louca, S. and M. W. Pennell. 2020b. A General and Efficient Algorithm for the Likelihood of Diversification and Discrete-Trait Evolutionary Models. Systematic Biology 69:545–556.

Maddison, W. P., P. E. Midford, and S. P. Otto. 2007. Estimating a Binary Character's Effect on Speciation and Extinction. Syst. Biol 56:701–710.

Magee, A. F. and S. Höhna. 2021. Impact of K-Pg Mass Extinction Event on Crocodylomorpha Inferred from Phylogeny of Extinct and Extant Taxa. bioRxiv Page 2021.01.14.426715.

Magnuson-Ford, K. and S. P. Otto. 2012. Linking the investigations of character evolution and species diversification. The American Naturalist 180:225–245.

Morlon, H. 2014. Phylogenetic approaches for studying diversification. Ecology Letters 17:508–525.

Morlon, H., T. L. Parsons, and J. B. Plotkin. 2011. Reconciling molecular phylogenies with the fossil record. Proceedings of the National Academy of Sciences of the United States of America 108:16327–16332.

Nee, S. 2006. Birth-Death Models in Macroevolution. Annual Review of Ecology, Evolution, and Systematics 37:1–17.

Nee, S., R. M. May, and P. H. Harvey. 1994. The reconstructed evolutionary process. Phil. Trans. R. Soc. B Page 7.

Ng, J. and S. D. Smith. 2014. How traits shape trees: New approaches for detecting character state-dependent lineage diversification. Journal of Evolutionary Biology 27:2035–2045.

Nielsen, R. 2002. Mapping Mutations on Phylogenies. Systematic Biology 51:729–739.

Pybus, O. G., A. Rambaut, and P. H. Harvey. 2000. An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies. Genetics 155:1429–1437.

Rabosky, D. L. and I. J. Lovette. 2008a. Density-dependent diversification in North American wood warblers. Proceedings of the Royal Society B: Biological Sciences 275:2363–2371.

Rabosky, D. L. and I. J. Lovette. 2008b. Explosive evoltuionary radiation: Decreasing specaition or increasing extinction through time? Evolution 62:1866–1875.

Rasmussen, D. A. and T. Stadler. 2019. Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models. eLife 8.

Raup, D. M. 1985. Mathematical models of cladogenesis. Paleobiology 11:42–52.

Romero-Severson, E. O., I. Bulla, and T. Leitner. 2016. Phylogenetically resolving epidemiologic linkage. Proceedings of the National Academy of Sciences 113:2690–2695.

Schluter, D. and M. W. Pennell. 2017. Speciation gradients and the distribution of biodiversity. Nature 546:48–55.

Silvestro, D., J. Schnitzler, L. H. Liow, A. Antonelli, and N. Salamin. 2014. Bayesian Estimation of Speciation and Extinction from Incomplete Fossil Occurrence Data. Systematic Biology 63:349–367.

Stadler, T. 2009. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. Journal of Theoretical Biology 261:58–66.

Stadler, T. 2010. Sampling-through-time in birth-death trees. Journal of Theoretical Biology 267:396–404.

Stadler, T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proceedings of the National Academy of Sciences of the United States of America 108:6187–6192.

Stadler, T. and S. Bonhoeffer. 2013. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. Philos. Trans. R. Soc. B 368:20120198.

Stadler, T., R. D. Kouyos, V. von Wyl, S. Yearly, and J. Böni. 2012. Estimating the Basic Reproductive Number from Viral Sequence Data. Mol. Biol. Evol. .

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

Stadler, T., T. G. Vaughan, A. Gavryushkin, S. Guindon, D. Kühnert, G. E. Leventhal, and A. J. Drummond. 2015. How well can the exponential-growth coalescent approximate constant-rate birth-death population dynamics? Proceedings. Biological Sciences 282:20150420.

Strimmer, K. and O. G. Pybus. 2001. Exploring the Demographic History of DNA Sequences Using the Generalized Skyline Plot. Molecular Biology and Evolution 18:2298–2305.

Vaughan, T. G., G. E. Leventhal, D. A. Rasmussen, A. J. Drummond, D. Welch, and T. Stadler. 2019. Estimating Epidemic Incidence and Prevalence from Genomic Data. Molecular Biology and Evolution 36:1804–1816.

Volz, E. M. 2012. Complex Population Dynamics and the Coalescent Under Neutrality. Genetics 190:187–201.

Volz, E. M., S. L. Kosakovsky Pond, M. J. Ward, A. J. Leigh Brown, and S. D. W. Frost. 2009. Phylodynamics of Infectious Disease Epidemics. Genetics 183:1421–1430.

Contents

I Single-type model **2**

S1 Relationships between existing models **2**

S1.1 Stadler 2009 2

S1.2 Stadler 2010 3

S1.3 Morlon *et al.* 2011 5

S1.4 Stadler *et al.* 2011 6

S1.5 Stadler *et al.* 2012 8

S1.6 Stadler *et al.* 2013 and Gavryushkina *et al.* 2014 9

S1.7 Heath *et al.* 2014 13

S2 Tables **15**

II The multi-type model **18**

S3 The multi-type diversification model **18**

S3.1 Derivation of $\mathcal{L}\left(\mathcal{T}, \mathcal{C} | \Theta_{MBDS}\right)$ 18

S3.2 Integrating over tree colourings 22

S4 Tables **29**

Part I

Single-type model

S1 Relationships between existing models

In the previous section, we proved that one could go from the most general case to specific sub-models by incorporating additional constraints to the parameters. In this section, we illustrate how to work in the other direction — that is, to start with the most assumptions of a particular sub-model and derive its likelihood function using the same five-step procedure used to derive the general BDS model in the Main Text. In addition to illustrating the utility of our mathematical technique, by deriving the likelihoods of previously developed models, we are able to unify a diverse and, occasionally opaque, literature using a common terminology, notation, and formulation.

S1.1 Stadler 2009

Here we re-derive the likelihood given by **Equation 2** in (Stadler, 2009). Note throughout all equation, corollary, and theorem references in other publications will be placed in bold face type.

- Step 1: Specify the model.
  - Constant rates: $\lambda(\tau) = \lambda$, $\mu(\tau) = \mu$.
  - Birth-death model with incomplete sampling at present day: $\psi(\tau) = 0$ and $\rho_0 \leq 1$.
  - Conditioning on there being exactly N_0 lineages at the present day given the time of origin, $\mathcal{S}_6$ and un-ordered birth events $\mathcal{S}_8 = (N_0 - 1)!$.
- Step 2: IVP for $g_e(\tau)$.

$$\begin{aligned} \frac{dg_e(\tau)}{d\tau} &= -(\lambda - \mu)g_e(\tau) + 2\lambda g_e(\tau)E(\tau) \\ g_e(\tau) &= \begin{cases} \lambda g_{e1}(s_e)g_{e2}(s_e) & \text{birth} \\ \rho_0 & \text{present day} \end{cases} \end{aligned}$$

- Step 3: IVP for $E(\tau)$.

$$\frac{dE(\tau)}{d\tau} = -(\lambda + \mu)E(\tau) + \lambda E(\tau)^2 + \mu \quad \text{where } E(0) = 1 - \rho_0.$$

  Given the constant rate assumption there exists a general solution for $E(\tau)$.

$$\begin{aligned} E(\tau) =& \frac{\lambda + \mu + c_1 \frac{\exp[-c_1 t](1-c_2)-(1+c_2)}{\exp[-c_1 t](1-c_2)+(1+c_2)}}{2\lambda} \\ c_1 =& \lambda - \mu \quad c_2 = \frac{\lambda - \mu - 2\lambda\rho_0}{\lambda - \mu} \end{aligned}$$

- Step 4: Derive $g_{stem}(T)$.

$$g_{stem}(T) = \rho_0^{N_0} \lambda^{N_0-1} \prod_{edges} \Psi(s_e, t_e).$$

- Step 5: Calculate $g_{stem}(T)$ wrt the critical time representation. Given the assumption of constant rates and no Poisson sampling the expression for $\Phi(\tau)$ simplifies to:

$$\Phi(\tau) = \frac{e^{-x(\lambda-\mu)}(\lambda-\mu)^2}{\left((\lambda(1-\rho_0)-\mu)e^{-x(\lambda-\mu)} + \lambda\rho_0\right)^2}.$$

  Hence we have:

$$g_{stem}(T) = \rho_0^{N_0} \lambda^{N_0-1} \prod_{i=1}^{N_0-1} \frac{e^{-x_i(\lambda-\mu)}(\lambda-\mu)^2}{\left((\lambda(1-\rho_0)-\mu)e^{-x_i(\lambda-\mu)} + \lambda\rho_0\right)^2}.$$

- Step 6: Impose conditioning. The likelihood is conditioned on having exactly N_0 lineages at the present day ($\mathcal{S}_6$) and a constant $\mathcal{S}_8 = (N_0-1)!$ as the birth events are left un-ordered.

$$\boxed{\mathcal{L}_{S09} = \frac{(N_0-1)!}{\hat{E}_{N_0}(T)} \rho_0^{N_0} \lambda^{N_0-1} \prod_{i=1}^{N_0-1} \frac{e^{-x_i(\lambda-\mu)}(\lambda-\mu)^2}{\left((\lambda(1-\rho_0)-\mu)e^{-x_i(\lambda-\mu)} + \lambda\rho_0\right)^2}.} \tag{S1}$$

S1.2 Stadler 2010

- Step 1: Specify the model.
  - Constant rates: $\lambda(\tau) = \lambda$, $\mu(\tau) = \mu$, $\psi(\tau) = \psi$.
  - No removal upon sampling, $r = 0$.
  - Multiple presented.
    - **Equation 3** (Stadler, 2010): No conditioning.
    - **Equation 4** (Stadler, 2010): Exactly N_0 extant sampled tips.
    - **Corollary 3.7** (Stadler, 2010): At least one extant tip conditioning on the time of origin.
    - **Equation 5** (Stadler, 2010): At least one extant tip conditioning on the t_{MRCA}.
    - **Equation 6** (Stadler, 2010): N_0 extant tips conditioning on the time the t_{MRCA}.
- Step 2: IVP for $g_e(\tau)$.

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

$$g_e(s_e) = \begin{cases} \lambda g_{e1}(s_e) g_{e2}(s_e) & \text{birth event giving rise to edges e1 and e2} \\ \psi g_e(s_e) & \text{ancestral sampling event} \\ \psi E(s_e) & \text{terminal sampling event} \\ \rho_0 & s_e = 0\text{, edge sampled at present day} \end{cases}$$

- Step 3: IVP for $E(\tau)$. Given the constant rate assumption there exists a general solution for $E(\tau)$.

$$\frac{dE(\tau)}{d\tau} = -(\lambda+\mu+\psi)\, E(\tau) + \lambda E(\tau)^2 + \mu \quad \text{where } E(0) = 1-\rho_0.$$

This is a Bernoulli differential equation and has a known solution as given by Stadler (Stadler, 2010).

$$E(t) = \frac{\lambda + \mu + c_1 \frac{\exp[-c_1 t](1-c_2)-(1+c_2)}{\exp[-c_1 t](1-c_2)+(1+c_2)}}{2\lambda}$$
$$c_1 = \left|\sqrt{(\lambda - \mu - \psi)^2 + 4\lambda\psi}\right| \quad c_2 = \frac{\lambda - \mu - 2\lambda\rho_0 - \psi}{c_1}.$$

- Step 4: Derive $g_{stem}(T)$.

$$g_{stem}(T) = \underbrace{\rho_0^{N_0}}_{\substack{\text{extant}\\\text{tips}}} \underbrace{\prod_{i=1}^{I} \lambda}_{\text{births}} \underbrace{\prod_{j=1}^{n} \psi E(y_j)}_{\substack{\text{extinct}\\\text{tips}}} \underbrace{\prod_{k=1}^{m} \psi}_{\substack{\text{ancestral}\\\text{samples}}} \underbrace{\prod_{e\in\mathcal{T}} \Psi(s_e, t_e)}_{\text{edges}}.$$

- Step 5: Calculate $g_{stem}(T)$ wrt the critical time representation. Given the assumption of constant rates and no Poisson sampling the expression for $\Phi(\tau)$ simplifies to:

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

Hence we have:

$$g_{stem}(T) = \underbrace{\Phi(T)}_{\text{root}} \underbrace{\rho_0^{N_0}}_{\substack{\text{extant}\\\text{tips}}} \underbrace{\prod_{i=1}^{I} \lambda\Phi(x_i)}_{\text{births}} \underbrace{\prod_{j=1}^{n} \frac{\psi}{\Phi(y_j)} E(y_j)}_{\substack{\text{extinct}\\\text{tips}}} \underbrace{\prod_{k=1}^{m} \psi}_{\substack{\text{ancestral}\\\text{samples}}}.$$

- Step 6: Impose conditioning. Imposing an arbitrary conditioning we have the following likelihood. Note that $I = N_0 + n - 1$.

$$\boxed{\mathcal{L}_{S10} = \mathcal{S}\Phi(T)\rho_0^{N_0}\lambda^{N_0+n-1}\psi^{n+m} \prod_{i=1}^{N_0+n-1} \Phi(x_i) \prod_{j=1}^{n} \frac{E(y_j)}{\Phi(y_j)}} \tag{S2}$$

  - **Equation 3** (Stadler, 2010): No conditioning. $\mathcal{S} = \mathcal{S}_0 = 1$.
  - **Equation 4** (Stadler, 2010): Exactly N_0 extant sampled tips. $\mathcal{S} = \mathcal{S}_6 = \frac{1}{\hat{E}_{N_0}(T)}$ where in the case of constant rates we have:

$$\hat{E}_{N_0}(T) = \hat{\phi}(T)\left(\frac{\mu}{\lambda}E(T)\right)^{N_0-1}.$$

    where $\hat{\phi}(\tau)$ equals $\phi(\tau)$ where $\psi = 0$.
  - **Corollary 3.7** (Stadler, 2010): At least one extant tip conditioning on the time of origin. $\mathcal{S} = \mathcal{S}_4 = \frac{1}{1-\hat{E}(T)}$ where $\hat{E}(T)$ is given by the solution for $E(\tau)$ above given that $\psi = 0$.
  - **Equation 5** (Stadler, 2010): At least one extant tip conditioning on the t_{MRCA}, $\mathcal{S} = \mathcal{S}_{5'} = \frac{\phi(x_1)}{\phi(T)(1-\hat{E}(x_1)^2}$.
  - **Equation 6** (Stadler, 2010): N_0 extant tips conditioning on the time the t_{MRCA}, $\mathcal{S} = \mathcal{S}_7 = \frac{\Phi(x_1)}{\Phi(t_{\text{or}})}\left(\sum_{i=1}^{N_0-1} \hat{E}_i(x_1)\hat{E}_{N_0-i}(x_1)\right)^{-1}$,

S1.3 Morlon *et al.* 2011

Here we derive **Equation 1** from (Morlon et al., 2011).

- Step 1: Specify the model.
  - Time variable rates.
  - Birth-death only $\psi(\tau) = 0$.
  - At least one extant sample $\mathcal{S}_2 = \mathcal{S}_4$.
- Step 2: IVP for $g_e(\tau)$.

$$\frac{dg_e(\tau)}{d\tau} = -\left(\lambda(\tau) + \mu(\tau)\right) g_e(\tau) + 2\lambda(\tau) g_e(\tau) E(\tau)$$

$$g_e(s_e) = \begin{cases} \lambda g_{e1}(s_e) g_{e2}(s_e) & \text{birth event giving rise to edges e1 and e2} \\ \rho_0 & s_e = 0\text{, edge sampled at present day} \end{cases}$$

- Step 3: IVP for $E(\tau)$.

$$\frac{dE(\tau)}{d\tau} = -\left(\lambda(\tau) + \mu(\tau)\right) E(\tau) + \lambda(\tau) E(\tau)^2 + \mu(\tau) \quad \text{where } E(0) = 1 - \rho_0.$$

  The general solution of this differential equation is given by:

$$E(\tau) = 1 - \frac{\rho_0 \exp\left[\int_0^\tau \left(\lambda(u) - \mu(u)\right) du\right]}{1 + \int_0^\tau \rho_0 \exp\left[\int_0^x \left(\lambda(u) - \mu(u)\right) du\right] dx},$$

  see **Equation 2** in (Morlon et al., 2011).

- Step 4: Derive $g_{stem}(T)$.

$$g_{stem}(T) = \underbrace{\rho_0^{N_0}}_{\substack{\text{extant}\\ \text{tips}}} \underbrace{\prod_{i=1}^{I} \lambda(x_i)}_{\text{births}} \underbrace{\prod_{e \in \mathcal{T}} \Psi(s_e, t_e)}_{\text{edges}},$$

  where the expression for $\Psi(s_e, t_e)$ is given by **Equation 3** in (Morlon et al., 2011).

- Step 5: Calculate $g_{stem}(T)$ wrt the critical time representation. Given $\Phi(\tau) = \Psi(0, \tau)$ from **Equation 3** for in (Morlon et al., 2011) we have:

$$\Phi(\tau) = \exp\left[\int_0^\tau \left(\lambda(\sigma) - \mu(\sigma)\right) d\sigma\right] \left[1 + \frac{\int_0^\tau \rho_0 \lambda(u) \exp\left[\int_0^u \left(\lambda(\sigma) - \mu(\sigma)\right) d\sigma\right] du}{1 + \rho_0}\right]^{-2}.$$

  Hence we have:

$$g_{stem}(T) = \underbrace{\Phi(T)}_{\text{root}} \underbrace{\rho_0^{N_0}}_{\substack{\text{extant}\\ \text{tips}}} \underbrace{\prod_{i=1}^{N_0 - 1} \lambda(x_i) \Phi(x_i)}_{\text{births}}.$$

  Note $I = N_0 - 1$.

- Step 6: Impose conditioning. The likelihood given by **Equation 1** in (Morlon et al., 2011) is conditioned on the existence of at least one sampled lineage, $\mathcal{S} = \mathcal{S}_3 = \frac{1}{1-E(t_{\text{or}})}$.

$$\boxed{\mathcal{L}_{M11} = \frac{\rho_0^{N_0}}{1-E(t_{\text{or}})}\Phi(T)\prod_{i=1}^{N_0-1}\lambda(x_i)\Phi(x_i)} \tag{S3}$$

S1.4 Stadler *et al.* 2011

Here we derive the likelihoods given by **Theorem 2.6** and **2.7** in (Stadler, 2011).

- Step 1: Specify the model.
  - piecewise-constant Poissonian rates. $\lambda(\tau) = \lambda_l$ and $\bar{\mu}(\tau) = \bar{\mu}_l$ if $t_l \leq \tau < t_{l+1} \quad l = 0, 2, ...L+1$ where $t_0 = 0$ and $t_{L+1} = T$.
  - No Poisson sampling, $\psi(\tau) = 0$.
  - Mass extinction events at times $t_l \quad l = 1, 2, ...L$ as specified above.

$$\begin{aligned}\mu(\tau) =& \bar{\mu} + \sum_{l}^{L} m_l \delta(\tau - t_l)\\ m_l =& -ln(1-\nu_l).\end{aligned}$$

  - **Theorem 2.6** (Stadler, 2011) imposes no additional conditioning $\mathcal{S} = \mathcal{S}_0$ whereas **Theorem 2.7** (Stadler, 2011) conditions on observing at least one descendent given the time of the most recent common ancestor $\mathcal{S} = \mathcal{S}_2 = \mathcal{S}_4$.

- Step 2: IVP for $g_e(\tau)$.

$$\begin{aligned}\frac{dg_e(\tau)}{d\tau} =& -(\lambda(\tau)+\mu(\tau))g_e(\tau) + 2\lambda(\tau)g_e(\tau)E(\tau)\\ g_e(s_e) =& \begin{cases}\lambda(s_e)g_{e1}(s_e)g_{e2}(s_e) & \text{birth event giving rise to edges e1 and e2}\\ \rho_0 & s_e = 0\text{, edge sampled at present day,}\end{cases}\end{aligned}$$

  where $\mu(\tau)$ is given above.

  In terms of the probability flow $g_e(\tau) = g_e(s_e)\Psi(s_e, \tau)$, where:

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

  Here $\mu(\tau)$ includes the mass extinction events.

- Step 3: IVP for $E(\tau)$.

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

Given the piecewise constant nature there is a known general solution. Let $E(\tau) = E_l(\tau)$ where $t_l < \tau \leq t_{l+1}$. Then define $E_{l-1}(t_l^-)$ as the solution up to but not including the mass extinction event at time t_l we have:

$$E_l(t_l) = E_{l-1}(t_l) = (1-\nu_l)E_{l-1}(t_l^-) + \nu_l,$$

where $E_l(\tau)$ is given by a solution similar to that in Equation (A1).

$$\begin{aligned} E_l(\tau) =& \frac{\lambda_l + \bar{\mu}_l}{2\lambda_l} + \frac{c_1}{2\lambda_l}\frac{e^{-c_1 t}(1-c_2) - (1+c_2)}{e^{-c_1 t}(1-c_2) + (1+c_2)} \\ c_1 =& \left|\sqrt{(\lambda_l - \bar{\mu}_l)^2}\right| \quad c_2 = -\frac{\lambda_l - \bar{\mu}_l - 2\lambda_l(1 - E_l(t_l))}{c_1}, \end{aligned}$$

where $E_0(t_0) = 1 - \rho_0$.

- Step 4: Derive $g_{stem}(T)$. The expression for $g_{stem}(T)$ is given by:

$$g_{stem}(T) = \underbrace{\rho_0^{N_0}}_{\text{extant tips}} \underbrace{\prod_{i=1}^{I} \lambda(x_i)}_{\text{births}} \underbrace{\prod_{e \in \mathcal{T}} \Psi(s_e, t_e)}_{\text{edges}},$$

where $\lambda(\tau)$ and $\Psi(s_e, \tau)$ are specified above.

- Step 5: The critical time representation. We once again define the sub-flow $\Phi_l(\tau)$ where $t_l < \tau \leq t_{l+1}$:

$$\begin{aligned} \Phi_l(\tau) =& \exp\left[\int_{t_l}^{\tau} 2\lambda(x)E(x) - (\lambda(x) + \bar{\mu}(x) + m_l\delta(x - t_l))\, dx\right] \\ =& \exp\left[\int_{t_l}^{\tau} 2\lambda(x)E(x) - (\lambda(x) + \bar{\mu}(x))\, dx\right](1 - \nu_l) \\ =& \bar{\Phi}_l(\tau)\,(1 - \nu_l)\,. \end{aligned}$$

Note $\nu_0 = 0$. The complete flow is, as given by Equation (A3).

$$\begin{aligned} \Phi(\tau) =& \bar{\Phi}_{L_\tau}(\tau) \prod_{l=1}^{L_\tau} \bar{\Phi}_{l-1}(t_l)\,(1 - \nu_l) \\ \Phi(t_k) =& \underbrace{\bar{\Phi}_k(t_k)}_{1} \prod_{l=1}^{k} \bar{\Phi}_{l-1}(t_l)(1 - \nu_l) = \prod_{l=1}^{k} \bar{\Phi}_{l-1}(t_l)\,(1 - \nu_l)\,, \end{aligned}$$

where L_τ is once again the largest index l such that $t_l < \tau$.

The critical time representation of g_{stem} then is:

$$g_{stem}(T) = \rho_0^{N_0} \bar{\Phi}_L(T) \prod_{l=1}^{L} \bar{\Phi}_{l-1}(t_l)(1 - \nu_l) \prod_{i=1}^{I} \left[\lambda(x_i)\bar{\Phi}_{L_{x_i}}(x_i) \prod_{l=1}^{L_{x_i}} \bar{\Phi}_{l-1}(t_l)(1 - \nu_l)\right].$$

Defining α_l as the number of observed birth events $\geq t_l$ we can rewrite the product:

$$\prod_{i=1}^{I}\prod_{l=1}^{L_{x_i}}\bar{\Phi}_{l-1}(t_l) = \prod_{l=1}^{L}\left(\bar{\Phi}_{l-1}(t_l)\right)^{\alpha_l}$$
$$\prod_{i=1}^{I}\prod_{l=1}^{L_{x_i}}(1-\nu_l) = \prod_{l=1}^{L}(1-\nu_l)^{\alpha_l}$$

Hence we have:

$$\begin{aligned} g_{stem}(T) =& \rho_0^{N_0}\bar{\Phi}_L(T)\prod_{l=1}^{L}\left(\bar{\Phi}_{l-1}(t_l)(1-\nu_l)\right)^{\alpha_l+1}\prod_{i=1}^{I}\left[\lambda(x_i)\bar{\Phi}_{L_{x_i}}(x_i)\right] \\ =& \rho_0^{N_0}\bar{\Phi}_L(T)\prod_{l=1}^{L}\left(\bar{\Phi}_{l-1}(t_l)(1-\nu_l)\right)^{n_l}\prod_{i=1}^{I}\left[\lambda(x_i)\bar{\Phi}_{L_{x_i}}(x_i)\right], \end{aligned}$$

where n_l is the number of lineages in the observed phylogeny at time t_l then $n_l = \alpha_l + 1$.

- Step 6: Likelihood conditioning.

  For **Theorem 2.6** (Stadler, 2011) the likelihood is given by:

$$\boxed{\mathcal{L}_{S11}(\mathcal{T}) = \rho_0^{N_0}\bar{\Phi}_L(T)\prod_{l=1}^{L}\left(\bar{\Phi}_{l-1}(t_l)(1-\nu_l)\right)^{n_l}\prod_{i=1}^{I}\left[\lambda(x_i)\bar{\Phi}_{L_{x_i}}(x_i)\right]}$$

  For **Theorem 2.7** (Stadler, 2011) the likelihood is given by:

$$\boxed{\mathcal{L}_{S11}(\mathcal{T}) = \frac{\rho_0^{N_0}}{(1-E(x_1))^2}\bar{\Phi}_{L_{x_1}}(x_1)\prod_{l=1}^{L}\left(\bar{\Phi}_{l-1}(t_l)(1-\nu_l)\right)^{n_l}\prod_{i=1}^{I}\left[\lambda(x_i)\bar{\Phi}_{L_{x_i}}(x_i)\right]}$$

S1.5 Stadler *et al.* 2012

Here we derive **Equation 1** in (Stadler et al., 2012).

- Step 1: Specify the model.
  - Constant birth, death, and sampling rates.
  - No present day sampling. All lineages removed upon sampling $r = 1$.
  - No conditioning.
- Step 2: IVP for $g_e(\tau)$.

$$\begin{aligned} \frac{dg_e(\tau)}{d\tau} &= -(\lambda+\mu+\psi)g_e(\tau) + 2\lambda g_e(\tau)E(\tau) \\ g_e(s_e) &= \begin{cases} \lambda(s_e)g_{e1}(s_e)g_{e2}(s_e) & \text{birth event giving rise to edges e1 and e2} \\ \psi(s_e) & \text{terminal sampling event} \end{cases} \end{aligned}$$

- Step 3: IVP for $E(\tau)$.

$$\frac{dE(\tau)}{d\tau} = -(\lambda+\mu+\psi)E(\tau) + \lambda E(\tau)^2 + \mu \quad \text{where } E(0) = 0.$$

The solution to this differential equation is given by:

$$E(t) = \frac{\lambda + \mu + c_1 \frac{\exp[-c_1 t](1-c_2)-(1+c_2)}{\exp[-c_1 t](1-c_2)+(1+c_2)}}{2\lambda}$$
$$c_1 = \left|\sqrt{(\lambda-\mu-\psi)^2 + 4\lambda\psi}\right| \quad c_2 = \frac{\lambda-\mu-\psi}{c_1}.$$

- Step 4: Derive $g_{stem}(T)$.

$$g_{stem}(T) = \underbrace{\prod_{i=1}^{I} \lambda}_{\text{births}} \underbrace{\prod_{j=1}^{n} \psi}_{\substack{\text{extinct}\\ \text{tips}}} \underbrace{\prod_{e\in\mathcal{T}} \Psi(s_e, t_e)}_{\text{edges}}$$
$$= \lambda^{n-1}\psi^n \prod_{e\in\mathcal{T}} \Psi(s_e, t_e),$$

where $I = n-1$.

- Step 5: Critical time representation. Given the assumption of constant rates and no Poisson sampling the expression for $\Phi(t)$ simplifies to:

$$\Phi(\tau) = \exp\left[\int_0^\tau 2\lambda E(x) - (\lambda+\mu+\psi)\, dx\right].$$

Hence we have:

$$g_{stem}(T) = \underbrace{\Phi(T)}_{\text{root}} \lambda^{n-1}\psi^n \underbrace{\prod_{i=1}^{n-1} \Phi(x_i)}_{\text{births}} \underbrace{\prod_{j=1}^{n} \frac{1}{\Phi(y_j)}}_{\substack{\text{extinct}\\ \text{tips}}}$$

- Step 6: Impose conditioning $\mathcal{S} = \mathcal{S}_0$.

$$\boxed{\mathcal{L}_{S12} = \lambda^{n-1}\psi^n \Phi(T) \prod_{i=1}^{n-1} \Phi(x_i) \prod_{j=1}^{n} \frac{1}{\Phi(y_j)}} \tag{S4}$$

S1.6 Stadler *et al.* 2013 and Gavryushkina *et al.* 2014

Here we derive the tree likelihood given by **Theorem 1** in (Stadler and Bonhoeffer, 2013) and due to their shared the likelihood given by **Equation 4** in (Gavryushkina et al., 2014).

- Step 1: Specify the model.
  - Piecewise constant Poissonian birth, death, and sampling rates.
    - define transition times $0 = t_0 < t_1 < ... < t_{L+1} = T$.

    * $\lambda(\tau) = \lambda_l t_l < \tau \leq t_{l+1}$.
    * $\mu(\tau) = \mu_l t_l < \tau \leq t_{l+1}$.
    * $\bar{\psi}(\tau) = \bar{\psi}_l t_l < \tau \leq t_{l+1}$.
  - Fossils/Ancestral sampling:
    * (Stadler and Bonhoeffer, 2013): No fossils $r = 1$.
    * (Gavryushkina et al., 2014): piecewise-constant rate $r(\tau) = r_l \, t_l < \tau \leq t_{l+1}$.
  - Concerted sampling attempts at each internal transition time t_l where $l = 1, 2...L$.
    * The probability of a lineage being sampled during the CSA at time t_l is ρ_l.
    * The resulting total sampling rate is given by:

$$\begin{aligned} \psi(\tau) =& \bar{\psi}(\tau) + \sum_{l=1}^{L} w_l \delta(\tau - t_l) \\ w_l =& - ln(1 - \rho_l). \end{aligned}$$

  - Conditioning:
    * Stadler et al. (Stadler and Bonhoeffer, 2013)—At least one sampled lineage, $\mathcal{S} = \mathcal{S}_2$.
    * Gavryushkina et al. (Gavryushkina et al., 2014)— At least one sample ($\mathcal{S}_2$) and a constant giving the number of un-oriented phylogenies $\mathcal{S} = \mathcal{S}_8$.

- Step 2: Derive IVP for $g_e(\tau)$. As in section 1 we have:

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

$$g_e(s_e) = \begin{cases} \lambda(s_e)g_{e1}(s_e)g_{e2}(s_e) & \text{birth event giving rise to edges e1 and e2} \\ (1 - r(s_e))\bar{\psi}(s_e)g_{e1}(s_e) & \text{Poisson ancestral sampling event} \\ \bar{\psi}(s_e)r(s_e) + \bar{\psi}(s_e)(1 - r(s_e))E(s_e) & \text{Poisson terminal sampling event} \\ (1 - r(t_l))\rho_l g_{e1}(t_l) & \text{ancestral sample at } t_l \\ \rho_l r(t_l) + \rho_l(1 - r(t_l))E(t_l) & \text{terminal sample at } t_l \\ \rho_0 & s_e = 0\text{, edge sampled at present day} \end{cases}$$

- Step 3: Derive IVP for $E(\tau)$. The IVP for $E(\tau)$ follows from Equation (8).

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

Given the piecewise constant nature there is a known general solution. Let $E(\tau) = E_l(\tau)$ where $t_l < \tau \leq t_{l+1}$. Then define $E_{l-1}(t_l^-)$ as the solution up to but not including the CSA at time t_l we have:

$$E_l(t_l) = E_{l-1}(t_l) = (1 - \rho_l)E_{l-1}(t_l^-),$$

where $E_l(\tau)$ is given by a solution similar to that in Equation (A1).

$$
\begin{aligned}
E_l(\tau) =& \frac{\lambda_l + \mu_l + \bar{\psi}_l}{2\lambda_l} + \frac{c_1}{2\lambda_l} \frac{e^{-c_1 t}(1-c_2) - (1+c_2)}{e^{-c_1 t}(1-c_2) + (1+c_2)} \\
c_1 =& \left| \sqrt{(\lambda_l - \mu_l - \bar{\psi}_l)^2 + 4\lambda_l + \bar{\psi}_l} \right| \quad c_2 = -\frac{\lambda_l - \mu_l - 2\lambda_l(1 - E_l(t_l)) - \bar{\psi}_l}{c_1},
\end{aligned}
$$

where $E_0(t_0) = 1 - \rho_0$.

- Step 4: Derive $g_{stem}(T)$. As in section 1 the edge representation of g_{stem} is given by:

$$
\begin{aligned}
g_{stem}(T) =& \rho_0^{N_0} \prod_{i=1}^{L} \lambda(x_i) \prod_{j=1}^{n} \psi(y_j) \left[(1 - r(y_j))E(y_j) + r(y_j)\right] \prod_{k=1}^{m} \psi(z_j)(1 - r(y_j)) \\
&\times \prod_{l=1}^{L} \rho_l \left[(1 - r_l)E(t_l) + r_l\right]^{N_l} \prod_{l=1}^{L} \left[\rho_l(1 - r_l)\right]^{M_l} \prod_{edges} \Psi(s_e, t_e),
\end{aligned}
$$

where $\Psi(s_e, t_e)$ is given by Equation (6).

- Step 5: Critical time representation. We once again define the sub-flow $\Phi_l(\tau)$ where $t_l < \tau \leq t_{l+1}$:

$$
\begin{aligned}
\Phi_l(\tau) =& \exp\left[\int_{t_l}^{\tau} 2\lambda(x)E(x) - \left(\lambda(x) + \mu(x) + \bar{\psi}(x) + w_l \delta(x - t_l)\right) dx\right] \\
=& \exp\left[\int_{t_l}^{\tau} 2\lambda(x)E(x) - \left(\lambda(x) + \mu(x) + \bar{\psi}(s)\right) dx\right] (1 - \rho_l) \\
=& \bar{\Phi}_l(\tau) \left(1 - \rho_l\right).
\end{aligned}
$$

The complete flow is, as given by Equation (A3).

$$
\begin{aligned}
\Phi(\tau) =& \bar{\Phi}_{L_\tau}(\tau) \prod_{l=1}^{L_\tau} \bar{\Phi}_{l-1}(t_l) \left(1 - \rho_l\right) \\
\Phi(t_k) =& \underbrace{\bar{\Phi}_k(t_k)}_{1} \prod_{l=1}^{k} \bar{\Phi}_{l-1}(t_l)(1 - \rho_l) = \prod_{l=1}^{k} \bar{\Phi}_{l-1}(t_l) \left(1 - \rho_l\right),
\end{aligned}
$$

where L_τ is once again the largest index l such that $t_l < \tau$. The critical time representation of g_{stem} from the previous stem gives the following.

$$
\begin{aligned}
g_{stem}(T) =& \rho_0^{N_0} \bar{\Phi}_L(T) \prod_{l=1}^{L} \bar{\Phi}_{l-1}(t_l)(1 - \rho_l) \prod_{i=1}^{I} \lambda(x_i) \bar{\Phi}_{L_{x_i}}(x_i) \prod_{i=1}^{I} \left[\prod_{l=1}^{L_{x_i}} \bar{\Phi}_{l-1}(t_l)(1 - \rho_l)\right] \\
&\times \prod_{j=1}^{n} \frac{\psi(y_j)}{\bar{\Phi}_{L_{y_j}}(y_j)} \left[(1 - r(y_j))E(y_j) + r(y_j)\right] \prod_{j=1}^{n} \left[\prod_{l=1}^{L_{y_j}} \frac{1}{\bar{\Phi}_{l-1}(t_l)(1 - \rho_l)}\right] \prod_{k=1}^{m} \psi(z_k)(1 - r(z_k)) \\
&\times \prod_{l=1}^{L} \left(\frac{\rho_l}{\bar{\Phi}_l(t_l)} \left[(1 - r_l)E(t_l) + r_l\right]\right)^{N_l} \prod_{l=1}^{L} \left[\prod_{j=1}^{l} \frac{1}{(1 - \rho_j)\bar{\Phi}_{j-1}(t_j)}^{N_l}\right] \prod_{l=1}^{L} \left[\rho_l(1 - r_l)\right]^{M_l}.
\end{aligned}
$$

We can then simplify the likelihood using the relationships similar to those given by Equation (A4).

$$\prod_{i=1}^{I}\prod_{l=1}^{L_{x_i}}\bar{\Phi}_{l-1}(t_l) = \prod_{l=1}^{L}\left[\bar{\Phi}_{l-1}(t_l)\right]^{\alpha_l}$$
$$\prod_{j=1}^{n}\prod_{l=1}^{L_{y_j}}\frac{1}{\bar{\Phi}_{l-1}(t_l)} = \prod_{l=1}^{L}\left[\bar{\Phi}_{l-1}(t_l)\right]^{-\sigma_l}.$$
$$\prod_{i=1}^{I}\prod_{l=1}^{L_{x_i}}(1-\rho_l) = \prod_{l=1}^{L}(1-\rho_l)^{\alpha_l}$$
$$\prod_{j=1}^{n}\prod_{l=1}^{L_{y_j}}\frac{1}{(1-\rho_l)} = \prod_{l=1}^{L}(1-\rho_l)^{-\sigma_l}$$
$$\prod_{l=1}^{L}\left[\prod_{j=1}^{l}\frac{1}{(1-\rho_j)^{N_l}}\right] = \prod_{l=1}^{L}\frac{1}{(1-\rho_l)^{\beta_l}},$$

where α_l is the number of birth events before time t_l and σ_l is the number of Poisson sampling events before time t_l and β_l is the number of lineages sampled during the CSAs up to and including at time t_l.

The resulting simplified likelihood is given by:

$$\begin{aligned} g_{stem}(T) =& \rho_0^{N_0}\bar{\Phi}_L(T)\prod_{l=1}^{L}\left[\bar{\Phi}_{l-1}(t_l)(1-\rho_l)\right]^{\alpha_l+1-\sigma_l-\beta_l} \\ &\times\prod_{j=1}^{n}\frac{\psi(y_j)}{\bar{\Phi}_{L_{y_j}}(y_j)}\left[(1-r(y_j))E(y_j)+r(y_j)\right]\prod_{k=1}^{m}\psi(z_k)(1-r(z_k)) \\ &\times\prod_{l=1}^{L}\left(\frac{\rho_l}{\bar{\Phi}_l(t_l)}\left[(1-r_l)E(t_l)+r_l\right]\right)^{N_l}\prod_{l=1}^{L}\left[\rho_l(1-r_l)\right]^{M_l}. \end{aligned}$$

- Step 6: Conditioned likelihood.
  - Stadler et al. (Stadler and Bonhoeffer, 2013) conditions on observing at least one sample. In addition the likelihood given by Stadler et al. assumes that all lineages are removed upon sampling $r_l = 1$ an assumption we will apply now. The likelihoods can be simplified by letting $n_l = \alpha_l + 1 - \sigma_l - \beta_l$ be the number of lineages that are alive immediately following the concerted sampling attempt at time t_l

$$\boxed{\mathcal{L}_{S13} = \frac{\rho_0^{N_0}}{1-E(T)}\bar{\Phi}_L(T)\prod_{l=1}^{L}\left[\bar{\Phi}_{l-1}(t_l)(1-\rho_l)\right]^{n_l}\prod_{j=1}^{n}\frac{\psi(y_j)}{\bar{\Phi}_{L_{y_j}}(y_j)}\prod_{l=1}^{L}\left(\frac{\rho_l}{\bar{\Phi}_l(t_l)}\right)^{N_l}.} \quad \text{(S5)}$$

  - Gavryushkina et al. (Gavryushkina et al., 2014) also condition on observing at least one lineage

as well as multiply by a constant giving the number of un-oriented trees.

$$\boxed{\begin{aligned}\mathcal{L}_{G14} =& \mathcal{S}_8 \frac{\rho_0^{N_0}}{1-E(T)} \bar{\Phi}_L(T) \prod_{l=1}^{L} \left[\bar{\Phi}_{l-1}(t_l)(1-\rho_l)\right]^{n_l} \\ &\times \prod_{j=1}^{n} \frac{\psi(y_j)}{\bar{\Phi}_{L_{y_j}}(y_j)} \left[(1-r(y_j))E(y_j) + r(y_j)\right] \prod_{k=1}^{m} \psi(z_k)(1-r(z_k)) \\ &\times \prod_{l=1}^{L} \left(\frac{\rho_l}{\bar{\Phi}_l(t_l)} \left[(1-r_l)E(t_l) + r_l\right]\right)^{N_l} \prod_{l=1}^{L} \left[\rho_l(1-r_l)\right]^{M_l}\end{aligned}} \tag{S6}$$

S1.7 Heath *et al.* 2014

- Step 1: Specify the model.
  - Constant rates: $\lambda(\tau) = \lambda$, $\mu(\tau) = \mu$, $\psi(\tau) = \psi$.
    - $*$ Here ψ denotes the sampling rate of fossils.
  - Birth-death model with fossils, $r = 0$.
  - Conditioned on observing ≥ 1 sample given t_{MRCA} ($\mathcal{S}_3$) and enumerated over all possible attachments of fossils (terminal sampling events before the present day ($\mathcal{S}_8$)).
- Step 2: Derive IVP for $g_e(\tau$.

$$\begin{aligned}\frac{dg_e(\tau)}{d\tau} =& -(\lambda+\mu+\psi)g_e(\tau) + 2\lambda g_e(\tau)E(\tau) \\ g_e(s_e) =& \begin{cases}\lambda g_{e1}(s_e)g_{e2}(s_e) & \text{birth event giving rise to edges e1 and e2} \\ \psi g_{e1}(s_e) & \text{ancestral sampling event} \\ \psi E(s_e) & \text{terminal sampling event} \\ \rho_0 & s_e = 0\text{, edge sampled at present day}\end{cases}\end{aligned}$$

- Step 3: Derive IVP for $E(\tau$.

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

  This has the same general solution as given above in section S1.2.

- Step 4: Derive $g_{stem}(T)$.

$$g_{stem}(T) = (\rho_0)^{N_0} \prod_{i=1}^{I} \lambda \prod_{j=1}^{n} \psi E(y_j) \prod_{k=1}^{m} \psi \prod_{edges} \Psi(s_e, t_e).$$

- Step 5: Critical time representation.

$$g_{stem}(T) = \Phi(T)\,(\rho_0)^{N_0} \prod_{i=1}^{I} \lambda\Phi(x_i) \prod_{j=1}^{n} \frac{\psi}{\Phi(y_j)} E(y_j) \prod_{k=1}^{m} \psi.$$

- Step 6: Conditioned likelihood. Conditioning on the probability that both daughter clades of the MRCA have at least one sampled extant descendant. $\mathcal{S}_5'$. The general solution to $\hat{E}(\tau)$ is known as given by the solution to $E(\tau)$ in section S1.1. Conditioning on all the possible attachments of the fossils (terminal samples before the present day). This requires multiplication by a constant $\mathcal{S}_8$. Let $\gamma(\tau)$ be the number of lineages alive at time τ. The the resulting number of fossil trees is.

$$\mathcal{S} = \prod_{i=1}^{n} 2\gamma(y_j) \prod_{k=1}^{m} \gamma(z_k). \tag{S7}$$

$$\boxed{\mathcal{L}_{H14} = \frac{\Phi(x_1)}{\left(1 - \hat{E}(x_1)\right)^2} (\rho_0)^{N_0} \prod_{i=1}^{I} \lambda\Phi(x_i) \prod_{j=1}^{n} \frac{2\gamma(y_j)\psi}{\Phi(y_j)} E(y_j) \prod_{k=1}^{m} \gamma(z_k)\psi} \tag{S8}$$

S2 Tables

General Birth-Death-Sampling Model Notation

Variable	Definition
$\lambda(\tau)$	The rate at which lineages speciate at time τ[a]. $\lambda : \mathbb{R} \to [0, \infty)$
$\mu(\tau)$	The rate at which lineages go extinct at time τ. $\mu : \mathbb{R} \to [0, \infty)$
$\psi(\tau)$	The rate at which lineages are sampled at time τ. $\psi : \mathbb{R} \to [0, \infty)$
$r(\tau)$	The probability that sampling is associated with host recovery/viral extinction at time τ. $r : \mathbb{R} \to [0, 1]$
ρ_0	The probability of sampling a lineage alive at the present day. $\rho_0 : \mathbb{R} \to (0, 1]$
T	The time of origin of the phylogeny/epidemic.
$\vec{x}$	The vector of I branching times. $\vec{x} = \{x_1, x_2, ..., x_i, ..., x_I\}$ [d]
$\vec{y}$	The vector of n internal (Poisson) sampling times without sampled descendants (terminal nodes). $\vec{y} = \{y_1, y_2, ..., y_j, ..., y_n\}$
$\vec{z}$	The vector of m internal (Poisson) sampling times of lineages with sampled descendants (sampled ancestors). $\vec{z} = \{z_1, z_2, ..., z_k, ..., z_m\}$
N_0	The number of lineage sampled at the present day.
$g_e(\tau)$	The likelihood density of observing a given phylogeny by the present day descending from a single edge e alive at age τ [e].
$E(\tau)$	The probability that a lineage alive at time τ leaves no sampled descendants in the phylogeny[f].
$\mathcal{S}$	The conditioning of the phylogeny.
$\Theta_{\text{BDS}} = \{\lambda(\tau), \mu(\tau), \psi(\tau), r(\tau), T, \rho_0\}$	
CSA [b], mass extinctions, and piecewise constant models	
L	The total number of past concerted sampling events.
L_t	The index of the time t_l at or after time t, i.e. the largest index such that $t_l \leq \tau$.
$\vec{t}$	Vector of times of CSAs in order of most recent to oldest $t_1 < t_2 < ... < t_L$. $\vec{t} = \{t_1, t_2, ..., t_l, ..., t_L\}$ [c]
$\vec{\rho}$	Vector of sampling probabilities of lineages sampled during each CSA. $\rho_l : \mathbb{R} \to (0, 1]$. $\vec{\rho} = \{\rho_0, \rho_1, ..., \rho_l, ..., \rho_L\}$
$\vec{r}$	The probability that a lineage that is sampled during each CSA is removed upon sampling. $r_l : \mathbb{R} \to []0, 1]$. $\vec{r} = \{r_1, r_2, ..., r_l, ..., r_L\}$
N_l	The number of tips sampled during the CSA at time t_l.
M_l	The number of ancestral samples during the CSA at time t_l.
n_l	The number of lineages that cross time t_l.
α_l	The number of observed birth events prior to time t_l ($x_i > t_l$).
σ_l (β_l)	The number of Poissonian (concerted) sampled lineages prior to or at time $\tau \geq t_l$.
$\Theta_{\text{CSA}} = \{\lambda(\tau), \mu(\tau), \psi(\tau), r(\tau), T, \vec{\tau}, \vec{\rho}, \vec{r}\}$	

Table S1: **model notation**.
[a]Throughout τ is measured in units of time before the present day ($\tau = 0$).
[b] CSA: Concerted Sampling Attempt.
[c] When used t_0 indicates the present day and $t_{L+1} = T$.
[d] When used x_1 indicates the time of the most recent common ancestor τ_{MRCA}
[e] The edge e spans time τ, $s_e \geq \tau \geq t_e$
[f] $\hat{E}(\tau)$ is the special case where $\psi(\tau) = 0$

Model	$\lambda(\tau), \mu(\tau), \psi(\tau)$	$r(\tau)$	ρ_0	ρ_l	r_l	ν_l	$\mathcal{S}$	Section
Stadler 2009	constant, $\psi = 0$	NA	> 0	0	NA	0	$\mathcal{S}_6\mathcal{S}_8$	S1.1
Stadler 2010	constant	0	> 0	0	NA	0	multiple	S1.2
Morlon et al. 2011	$\psi = 0$	NA	> 0	0	NA	0	$\mathcal{S}_1 = \mathcal{S}_2$	S1.3
Stadler 2011	piecewise, $\psi = 0$	NA	> 0	0	NA	≥ 0	$\mathcal{S}_3 = \mathcal{S}_5$	S1.4
Stadler et al. 2012	constant	1	0	0	NA	0	$\mathcal{S}_1$	S1.5
Stadler and Bonhoeffer 2013	piecewise	1	> 0	≥ 0	1	0	$\mathcal{S}_1$	S1.6
Gavryushkina et al. 2014	piecewise	piecewise	> 0	≥ 0	≥ 0	0	$\mathcal{S}_1 \times \mathcal{S}_7$	S1.6
Kühnert et al. 2014	stochastic SIR	constant	0	0	NA	0	?[a]	
Heath et al. 2014	constant	0	> 0	0	NA	0	$\mathcal{S}_6$	S1.7

Table S2: **Relationship of single-type sub-models to the general BDS model**.
[a] No specific conditioning mentioned.

Condition	Description	Examples
$\mathcal{S}_0 = 1$	No conditioning	**Eq.3**(Stadler, 2010), **Eq.1**(Stadler et al., 2012) & **Thrm. 2.6**(Stadler, 2011)
$\mathcal{S}_1 = \frac{\Phi(x_1)}{\Phi(T)}$	Likelihood given the t_{MRCA} rather than on the time of origin.	
$\mathcal{S}_2 = \frac{1}{1-E(T)}$	At least one sampled descendent either at or before the present day (given T).	**Eq.2**Morlon et al. (2011), **Thrm.1**Stadler and Bonhoeffer (2013)
$\mathcal{S}_3 = \frac{\Phi(x_1)}{\Phi(T)(1-E(x_1))^2}$	At least one sampled descendent at or before the present day given $\tau_{\text{MRCA}} = x_1$	(Heath et al., 2014)
$\mathcal{S}_4 = \frac{1}{1-\hat{E}(T)}$	At least one *extant* sampled lineage (given T).	**Cor.3.7** (Stadler, 2010) & **Thrm.2.7** (Stadler, 2011)
$\mathcal{S}_5 = \frac{\Phi(x_1)}{\Phi(T)2(1-\hat{E}(x_1))(1-E(x_1))}$	At least one *extant* sampled lineage given $\tau_{\text{MRCA}} = x_1$	
$\mathcal{S}_5' = \frac{\Phi(x_1)}{\Phi(T)(1-\hat{E}(x_1))^2}$	Both daughters of MRCA have at least one *extant* sampled lineage	**Eq. 5** Stadler (2010)
$\mathcal{S}_6 = \frac{1}{\hat{E}_{N_0}(T)}$	Exactly N_0 *extant* sampled lineages (given T).	**Eq.2**Stadler (2009), **Eq.4**(Stadler, 2010) &**Cor.3.6** Stadler (2010)
$\mathcal{S}_7 = \frac{\Phi(x_1)}{\Phi(t_{\text{or}})}\left(\sum_{i=1}^{N_0-1}\hat{E}_i(x_1)\hat{E}_{N_0-i}(x_1)\right)^{-1}$	Exactly N_0 *extant* sampled lineages given t_{MRCA}.	**Eq.6**Stadler (2010)
$\mathcal{S}_8 = \text{constant}$	Multiply by a constant	Stadler (2009) & Gavryushkina et al. (2014) [a] & Heath et al. (2014)

Table S3: **Alternative conditioning of tree likelihood**.
[a] See Gavryushkina et al. (2013) for algorithms for enumerating trees.

Part II

The multi-type model

S3 The multi-type diversification model

Here we consider the diversification of lineages of A discrete types. Lineages of type $a \in \{1, 2, ...A\}$ speciate/give birth to lineages of type $b \in \{1, 2, ...A\}$ at the time-variable rate $\lambda_{a,b}(\tau)$ at time τ before the present day. As above we will use τ to denote time moving backward from the present day ($\tau = 0$) to the origin of the phylogeny $\tau = T$. When $a = b$ speciation occurs without cladogenetic change, when $a \neq b$ specition is coincident with state change. In addition to the lineages changing state at birth events, lineages can mutate anagenetically from state a to state b at rate $\gamma_{a,b}(\tau)$. Lineages of type a alive at time τ, go extinct/die at rate $\mu_a(\tau)$. For $\tau > 0$ lineages are sampled at rate $\psi_a(\tau)$. Upon sampling a lineage may be removed from the population (e.g., sampling is coincident with treatment) with the state-dependent probability $r_a(\tau)$. Finally, all lineages alive at the present day are sampled with a state-dependent probability ρ_a. Model notation is summarized in Table S4.

As with the single-type model, the result of the mutli-type diversification process is a *full* and a *sampled* tree. Now however the tree is characterized by its topology $\mathcal{T}$ and the *colouring* of the tree $\mathcal{C}$ denoting the states of each lineage through time. Below we first derive an expression for the likelihood of a given coloured tree, $\mathcal{L}\left(\mathcal{T}, \mathcal{C}|\Theta_{MBDS}\right)$. However, as plausibly attainable data consists of knowledge at only some or all the sampled ancestral nodes and/or tips, we must then integrate the likelihood $\mathcal{L}\left(\mathcal{T}, \mathcal{C}|\Theta_{MBDS}\right)$ over all possible tree colourings consistent with the observed data.

S3.1 Derivation of $\mathcal{L}\left(\mathcal{T}, \mathcal{C}|\Theta_{MBDS}\right)$

We derive the likelihood using the steps used above for the single-type birth death sampling model. As with the single-type model for the likelihood calculation be begin by representing the phylogeny as a series of edges. However, as we are now referring to the coloured phylogeny we consider the set of all coloured edges, with each edge being a a segment of the phylogeny that is all of one colour beginning and ending at birth, sampling, or mutation events. Specifically, moving backward in time towards the tree origin let edge e start at time s_e before the present at a birth event, sampling (ancestral or tip), or mutation event, and continue toward the tree origin until time τ_e ending at either a birth event, ancestral sampling event, mutation event or at the tree origin.

As with the single-type model the tree likelihood depends on two different functions. First, $g_{e,a}(\tau)$ is the probability that an edge e with state a alive at time τ (hence $t_e > \tau > s_e$) gives rise to the subsequently observed phylogeny between τ and the present day. Second, $E_a(\tau)$ is the probability a lineage of type a alive at time τ has no sampled descendants between τ and the present day. We begin below by first deriving the initial value problems for these two functions.

Step 1: Derive the Initial Value Problem for $g_{e,a}(\tau)$.

To simplify notation we define the total birth and mutation rates of a given type $\Lambda_a = \sum_b \lambda_{a,b}$ and $\Gamma_a = \sum_b \lambda_{a,b}$ and the relative probability a birth or mutation event was of a given type, $p_{\lambda_{a,b}} = \frac{\lambda_{a,b}}{\Lambda_a}$ and $p_{\gamma_{a,b}} = \frac{\gamma_{a,b}}{\Gamma_a}$. For some small amount of time $\Delta\tau$ the recursion equation for $g_{e,a}(\tau)$ is given by:

$$
\begin{aligned}
g_{e,a}(\tau+\Delta\tau) \approx & \underbrace{\left((1-\Lambda_a\Delta\tau)(1-\mu_a\Delta\tau)(1-\psi_a\Delta\tau)(1-\Gamma_a\Delta\tau)\right) g_{e,a}(\tau)}_{\text{No Event}} \\
& + \underbrace{\left((\Lambda_a\Delta\tau)(1-\mu_a\Delta\tau)(1-\psi_a\Delta\tau)(1-\Gamma_a\Delta\tau)\right) \times \sum_b^A p_{\lambda_{a,b}} \iota_{a,b} g_{e,a}(\tau) E_b(\tau)}_{\text{Birth Events}} \\
& + \underbrace{\left((1-\Lambda_a\Delta\tau)(\mu_a\Delta\tau)(1-\psi_a\Delta\tau)(1-\Gamma_a\Delta\tau)\right) \times 0}_{\text{Death Event}} \\
& + \underbrace{\left((1-\Lambda_a\Delta\tau)(1-\mu_a\Delta\tau)(\psi_a\Delta\tau)(1-\Gamma_a\Delta\tau)\right) \times 0}_{\text{Sampling Event}} \\
& + \underbrace{\left((1-\Lambda_a\Delta\tau)(1-\mu_a\Delta\tau)(1-\psi_a\Delta\tau)(\Gamma_a\Delta\tau)\right) \times 0}_{\text{Mutation Events}} + \mathcal{O}(\Delta\tau^2)
\end{aligned}
\tag{S9}
$$

where $\iota_{a,b}$ is an indicator variable that has value 2 when $a = b$ and 1 otherwise. This recursion equation uses the fact that edges are all in state with mutation events resulting in the end of an edge and the origin of a daughter edge with a different state. This is reflected by the initial conditions of $g_{e,a}(\tau)$ as given below in equation (S11).

As above we can then use the definition of a derivative to obtain the corresponding differential equation

$$
\frac{dg_{e,a}(\tau)}{d\tau} = -\left(\Lambda_a + \mu_a + \psi_a + \Gamma_a\right) g_{e,a}(\tau) + \sum_b \iota_{a,b}\lambda_{a,b} g_{e,a}(\tau) E_b(\tau) \tag{S10}
$$

The initial conditions for $g_{e,a}$ at the start of the edge, time s_e, depend on the event that occurs in the sampled tree at that time. Specifically there are five types of events in the observed tree. 1) *Observed birth events* where a lineage of type a speciates producing a new lineage of type b. 2) *Ancestral sampling events* where a lineage of type a is sampled, is not removed from the population, and then has subsequently observed descendants. 3) *Terminal sampling events* where a lineage of type a is sampled and then either is immediately removed from the population or remains in the population but has no sampled descendants. 4) *Transition events* where a lineage of type a transitions to a lineage of type b. There are two possible ways a transition event can occur. First, a lineage can switch states due to a anagenetic *mutation event* or second an apparent transition can arise due to a birth event with cladogenetic state change followed by subsequent extinction of the parental lineage. We refer to this second form of transition event as a *hidden birth event*.

$$
g_{e,a}(s_e) = \begin{cases}
\lambda_{a,b}(s_e) g_{e_1,a}(s_e) g_{e_2,b}(s_e) & \text{birth event } a \to a+b \\
(1-r_a(s_e))\psi_a(s_e) g_{e_1,a}(s_e) & \text{ancestral sampling event} \\
r_a(s_e)\psi_a(s_e) + (1-r_a(s_e))\psi_a(s_e) E_a(s_e) & \text{terminal sampling event} \\
\left(\gamma_{a,b}(s_e) + \lambda_{a,b} E_a\right) g_{e_1,b}(s_e) & \text{mutation/hidden birth event } a \to b \\
\rho_a & \text{sampled at present day}
\end{cases}
\tag{S11}
$$

Importantly, as the ODES in equation (S10) are linear the corresponding IVP has the following general

solution given the initial conditions above.

$$\begin{aligned} g_{e,a}(\tau) =& g_{e,a}(s_e)\Psi_a(s_e,\tau), \quad \text{where} \\ \Psi_a(s_e,\tau) =& exp\left[\int_{s_e}^{\tau} -\left(\Lambda_a(x)+\mu_a(x)+\psi_a(x)+\Gamma_a(x)\right)+\sum_b \iota_{a,b}\lambda_{a,b}(x)E_b(x)dx\right]. \end{aligned} \tag{S12}$$

The IVP for the function $E_a(\tau)$ can be derived in an analogous manner.

Step 2: Derive the Initial Value Problem for $E_a(\tau)$ Once again we begin by deriving a recursion equation for the change in $E_a(\tau)$ over from time τ to $\tau+\Delta\tau$.

$$\begin{aligned} E_a(\tau+\Delta\tau) \approx& \underbrace{\left((1-\Lambda_a\Delta\tau)(1-\mu_a\Delta\tau)(1-\psi_a\Delta\tau)(1-\Gamma_a\Delta\tau)\right)E_a(\tau)}_{\text{No Event}} \\ &+\underbrace{\left((\Lambda_a\Delta\tau)(1-\mu_a\Delta\tau)(1-\psi_a\Delta\tau)(1-\Gamma_a\Delta\tau)\right)\times\sum_b^A p_{\lambda_{a,b}}E_a(\tau)E_b(\tau)}_{\text{Birth Events}} \\ &+\underbrace{\left((1-\Lambda_a\Delta\tau)(\mu_a\Delta\tau)(1-\psi_a\Delta\tau)(1-\Gamma_a\Delta\tau)\right)\times 1}_{\text{Death Event}} \\ &+\underbrace{\left((1-\Lambda_a\Delta\tau)(1-\mu_a\Delta\tau)(\psi_a\Delta\tau)(1-\Gamma_a\Delta\tau)\right)\times 0}_{\text{Sampling Event}} \\ &+\underbrace{\left((1-\Lambda_a\Delta\tau)(1-\mu_a\Delta\tau)(1-\psi_a\Delta\tau)(\Gamma_a\Delta\tau)\right)\times\sum_b p_{\gamma_{a,b}}E_b(\tau)}_{\text{Mutation Events}}+\mathcal{O}(\Delta\tau^2). \end{aligned} \tag{S13}$$

Using the definition of a derivative we have:

$$\frac{dE_a(\tau)}{d\tau} = -\left(\Lambda_a+\mu_a+\psi_a+\Gamma_a\right)E_a(\tau)+\sum_b \lambda_{a,b}E_a(\tau)E_b(\tau)+\mu_a+\sum_b \gamma_{a,b}E_b(\tau), \tag{S14}$$

with initial conditions given at the present day by:

$$E_a(0) = 1-\rho_a \tag{S15}$$

Step 3: Derive Expression for $g_{stem}(T)$ As in equation (9) for the single-type model, the tree likelihood is given by the value of the function g of the stem edge evaluated at the time of the origin of the phylogeny, T. From the solution to the initial value problem given by equation (S12) and the initial conditions (S11) we can write $g_{stem}(T)$ as a product over the events in the tree and all the probability flow Ψ of the edges in-between. To do so let $\vec{x}_{a,b}$ be a vector of length $I_{a,b}$, giving the time before the present day of all the observed birth events where lineages of type a give rise to lineages of type b. Let $\vec{y}_a$ be a vector of length J_a giving the sampling time of tips of type a and $\vec{z}_a$ a vector of length K_a the sampling type of sampled ancestors of type a. Finally let $\vec{w}_{a,b}$ be the time of observed transition events where a edge of type a transitions becomes an edge of type b. Once again, transition events can arise due to both direct mutation and hidden birth events. The resulting expression for the likelihood $g_{stem}(T)$ is given by:

$$
\begin{aligned}
g_{\textit{stem}}(T) = \prod_{a=1}^{A} \Bigg[& \underbrace{\rho_a^{N_a}}_{\text{present day sampling}} \times \underbrace{\prod_{b=1}^{A} \prod_{i=1}^{I_{a,b}} \lambda_{a,b}(x_{a,b,i})}_{\text{birth events}} \\
& \times \underbrace{\prod_{j=1}^{J_a} \psi_a(y_{a,j}) \left(1 - r_a(y_{a,j})\right) E_a(y_{a,j}) + \psi_a(y_{a,j}) r_a(y_{a,j})}_{\text{terminal samples}} \times \underbrace{\prod_{k=1}^{K_a} \psi_a(z_{a,k})(1 - r_a(z_{a,k}))}_{\text{ancestral samples}} \\
& \times \underbrace{\prod_{l=1}^{L_{a,b}} \left[\gamma_{a,b}(w_{a,b,l}) + \lambda_{a,b}(w_{a,b,l}) E_a(w_{a,b,l})\right]}_{\text{transitions}} \prod_{\text{edges of type } a} \Psi_a(s_e, \tau_e) \Bigg]
\end{aligned}
\tag{S16}
$$

Step 4: Rewrite $g_{\textit{stem}}(T)$ in Terms of Critical Times

Rather than enumerate Ψ over the edges of the phylogeny we can rewrite equation (S16) in terms of only the critical times $\vec{x}, \vec{y}, \vec{z}, \vec{w}$ Written in this form the likelihood also depends on c^* the colour of the phylogeny at the origin. To do so we define $\Phi_a(\tau) = \Psi_a(0, \tau)$ and rewrite the probability flow $\Psi_a(s_e, \tau)$ as a ratio:

$$
\Psi_a(s_e, \tau) = \frac{\Phi_a(\tau)}{\Phi_a(s_e)} \tag{S17}
$$

Substitution into equation (S16).

$$
\begin{aligned}
g_{\textit{stem},c^*}(T) = & \underbrace{\left[\prod_{a=1}^{A} \rho_a^{N_a}\right]}_{\text{present day sampling}} \times \underbrace{[\Phi_{c^*}(T)]}_{\text{stem}} \times \underbrace{\left[\prod_{a=1}^{A} \prod_{b=1}^{A} \prod_{i=1}^{I_{a,b}} \lambda_{a,b}(x_{a,b,i}) \frac{\cancel{\Phi_a(x_{a,b,i})} \Phi_b(x_{a,b,i})}{\cancel{\Phi_a(x_{a,b,i})}}\right]}_{\text{birth events}} \\
& \times \underbrace{\left[\prod_{a=1}^{A} \prod_{j=1}^{J_a} \left[\psi_a(y_{a,j})(1 - r(y_{a,j})) E_a(y_{a,j}) + \psi_a(y_{a,j}) r_a(y_{a,j})\right] \frac{1}{\Phi_a(y_{a,j})}\right]}_{\text{terminal sampling events}} \\
& \times \underbrace{\left[\prod_{a=1}^{A} \prod_{k=1}^{K_a} \psi_a(z_{a,k})(1 - r(z_{a,k})) \frac{\cancel{\Phi_a(z_{a,k})}}{\cancel{\Phi_a(z_{a,k})}}\right]}_{\text{ancestral sampling events}} \\
& \times \underbrace{\left[\prod_{a=1}^{A} \prod_{l=1}^{L_{a,b}} \left[\gamma_{a,b}(w_{a,b,l}) + \lambda_{a,b}(w_{a,b,l}) E_a(w_{a,b,l})\right] \frac{\Phi_a(w_{a,b,l})}{\Phi_b(w_{a,b,l})}\right]}_{\text{transition sampling events}}
\end{aligned}
\tag{S18}
$$

Step 5: Condition the Likelihood

We conclude, as in the single type model by multiplying the likelihood by including a general form of

conditioning $\mathcal{S}$.

$$
\begin{aligned}
\mathcal{L}(\mathcal{T}, \mathcal{C}|\Theta_{\text{MBDS}}) =& \mathcal{S} \times \left[\prod_{a=1}^{A} \rho_a^{N_a}\right] \times [\Phi_{c^*}(T)] \times \left[\prod_{a=1}^{A}\prod_{b=1}^{A}\prod_{i=1}^{I_{a,b}} \lambda_{a,b}(x_{a,b,i})\Phi_b(x_{a,b,i})\right] \\
&\times \left[\prod_{a=1}^{A}\prod_{j=1}^{J_a} [\psi_a(y_{a,j})(1-r(y_{a,j}))E_a(y_{a,j}) + \psi_a(y_{a,j})r_a(y_{a,j})] \frac{1}{\Phi_a(y_{a,j})}\right] \\
&\times \left[\prod_{a=1}^{A}\prod_{k=1}^{K_a} \psi_a(z_{a,k})(1-r(z_{a,k}))\right] \\
&\times \left[\prod_{a=1}^{A}\prod_{b\neq a}^{A}\prod_{l=1}^{L_{a,b}} [\gamma_{a,b}(w_{a,b,l}) + \lambda_{a,b}(w_{a,b,l})E_a(w_{a,b,l})] \frac{\Phi_a(w_{a,b,l})}{\Phi_b(w_{a,b,l})}\right]
\end{aligned}
\tag{S19}
$$

Due to the combination of cladogenetic change and different possibility of hidden birth events due to extinction and sampling, there are six different types of birth events included in the likelihood of the sampled tree. The following diagram summarizes these different birth events and how each is included in the equation (S19).

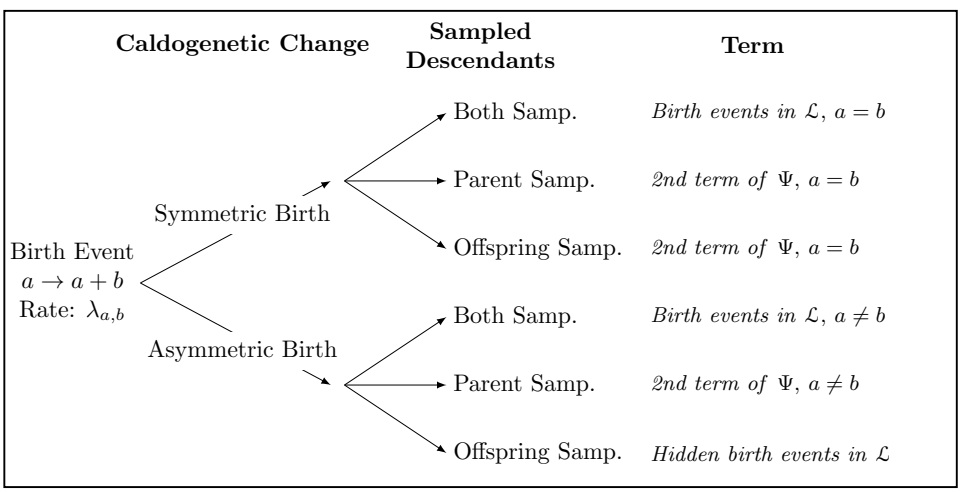

Figure S1: The six possible birth events and how each in included in the likelihood given by equation (S19).

S3.2 Integrating over tree colourings

The likelihood expression given by equation (S19) requires knowledge of when and how (e.g, through cladogenetic or anagenetic change) lineages have changed state, information which it is not plausible to obtain. Instead the available data is likely to include either only knowledge of the sampled tip states, which we denote as $\mathcal{C}_y$ for the colouring at the points in $\vec{y}$, and potentially the known states of sampled ancestors, $\mathcal{C}_z$. Note that the $\mathcal{C}_y$ and $\mathcal{C}_z$ need not include information for all sampled tips or sampled ancestors respectively. Calculating the corresponding likelihood of of the tree topology and the observed data can be calculating by

computing the following integral:

$$\mathcal{L}\left(\mathcal{T}, \mathcal{C}_y, \mathcal{C}_z | \Theta_{\text{MBDS}}\right) = \int_{\Omega_{\mathcal{C}|\mathcal{C}_y\mathcal{C}_z}} \mathcal{L}\left(\mathcal{T}, \mathcal{C} | \Theta_{\text{MBDS}}\right) Pr\left(\mathcal{C} | \mathcal{C}_y, \mathcal{C}_z \Theta_{\text{MBDS}}\right) d\mathcal{C} \tag{S20}$$

where $\Omega_{\mathcal{C}|\mathcal{C}_y\mathcal{C}_z}$ is the space of all possible tree colourings that are consistent with the observed states at the sampled tips and sampled ancestors. Here $Pr\left(\mathcal{C}|\mathcal{C}_y, \mathcal{C}_z\Theta_{\text{MBDS}}\right)$ is the probability of observing a given tree colouring given the observed states. Numerical calculation of this integral can be obtained via Monte Carlo through the use of a stochastic mapping algorithms (Nielsen, 2002) as exemplified below.

The BiSSE-ness model with time-variable rates

Here we examine one particular sub-model of the more general multi-type model which exemplifies many of the complexities of diversification with multiple states. Specifically we extend the BiSSEness/ClaSSE model (Magnuson-Ford and Otto, 2012; Goldberg and Igić, 2012) to allow for time-variable model rates. See *Mathematica* notebook for reference

The diversification model

The multi-type model above allows the birth, death, and sampling rates to vary arbitrarily through time. In order to specify realistic time-variable rates we begin by specifying a mechanistic model for the diversification process. Specifically we model the diversification dynamics with density-dependence where birth and death rates vary respect to a model of deterministic logistic growth with mutation. The deterministic number of lineages of type a through time, $n_a(t)$, is given by the solution to the following initial value problem:

$$\begin{aligned}\frac{dn_a(t)}{dt} =& n_a\left(b_{0,a} + b_{1,a}n_a\right)\left(1 - p_a\right) + n_{a'}\left(b_{0,a'} + b_{1,a'}n_{a'}p_{a'}\right)\\ & n_a\left(d_{0,a} + d_{1,a}n_a + d_{2,a}n_{a'}\right) - n_a\omega_a + n_{a'}\omega_{a'}\\ n_a(0) =& \pi(a),\end{aligned} \tag{S21}$$

where p_a is the probability that a birth event with a type a parental lineage undergoes cladogenetic change and π_a is the prior probability the tree is initially in state a. The mutation/transition rate from state a to a' is constant, occurring at a per-capita rate ω_a. Under this mechanistic model the diversification model rates are given by:

$$\begin{aligned}\lambda_{a,b} =& \begin{cases}\left(b_{0,a} + b_{1,a}n_a\right)\left(1 - p_a\right) & if\ a = b\\ \left(b_{0,a} + b_{1,a}n_a\right)p_a & if\ a \neq b\end{cases}\\ \mu_a =& \left(d_{0,a} + d_{1,a}n_a + d_{2,a}n_{a'}\right)\\ \gamma_{a,b} =& \omega_a.\end{aligned} \tag{S22}$$

Lineages of type a are sampled at the present day with probability ρ_a.

The BiSSEness likelihood with time-variable rates

The likelihood for the tree given the observed tip states can be derived in a similar manner as in (Magnuson-Ford and Otto, 2012), allowing the birth and death rates to vary through time.

First we derive the initial value problem for $D_{e,a}(\tau)$, the probability an edge e in the tree topology $\mathcal{T}$ in state a at time τ gives rise to the subsequently observed phylogeny. Note that in contrast to the function $g_{e,a}(\tau)$, the edge e here refers to an edge in the tree topology not in a coloured tree and hence can change state. The edge here begins at either the origin of the phylogeny or a birth event and ends at a birth event or at the tip in the phylogeny.

We begin by deriving the recursion equation for D between time τ and $\tau + \Delta\tau$.

$$\begin{aligned} D_{e,a}(\tau+\Delta\tau) \approx & \underbrace{\left((1-\Lambda_a\Delta\tau)(1-\mu_a\Delta\tau)(1-\psi_a\Delta\tau)(1-\Gamma_a\Delta\tau)\right) D_{e,a}(\tau)}_{\text{No Event}} \\ & + \underbrace{\left((\Lambda_a\Delta\tau)(1-\mu_a\Delta\tau)(1-\psi_a\Delta\tau)(1-\Gamma_a\Delta\tau)\right) \times \sum_b^A p_{\lambda_{a,b}} \iota_{a,b} D_{e,a}(\tau) E_b(\tau)}_{\text{Birth Events}} \\ & + \underbrace{\left((1-\Lambda_a\Delta\tau)(\mu_a\Delta\tau)(1-\psi_a\Delta\tau)(1-\Gamma_a\Delta\tau)\right) \times 0}_{\text{Death Event}} \\ & + \underbrace{\left((1-\Lambda_a\Delta\tau)(1-\mu_a\Delta\tau)(\psi_a\Delta\tau)(1-\Gamma_a\Delta\tau)\right) \times 0}_{\text{Sampling Event}} \\ & + \underbrace{\left((1-\Lambda_a\Delta\tau)(1-\mu_a\Delta\tau)(1-\psi_a\Delta\tau)(\Gamma_a\Delta\tau)\right) \times D_{e,a'}}_{\text{Mutation Events}} + \mathcal{O}(\Delta\tau^2). \end{aligned} \tag{S23}$$

This equation is similar to that for $g_{e,a}$ except for the mutation term which is no-longer multiplied by 0, but rather by $D_{e,a'}$. As a result the corresponding IVP is a system of coupled differential equations and no-longer has a known general solution. Similarly, as the type of birth event occurring at each internal node is now unknown, we must sum over the probability of all possible events at these nodes.

$$\begin{aligned} \frac{dD_{e,a}(\tau)}{d\tau} &= -\left(\Lambda_a + \mu_a + \psi_a + \Gamma_a\right) D_{e,a}(\tau) + \sum_b \iota_{a,b}\lambda_{a,b} D_{e,a}(\tau) E_b(\tau) + \gamma_{a,b} D_{e,b}(\tau) \\ D_{e,a}(s_e) &= \begin{cases} \sum_b \frac{\lambda_{a,b}(s_e)}{2}\left(D_{e_1,a}(s_e)D_{e_2,b}(s_e) + D_{e_1,b}(s_e)D_{e_2,a}(s_e)\right) & \text{birth event} \\ \rho_a & \text{sampled at present day} \end{cases} \end{aligned} \tag{S24}$$

The initial value problem for $E_a(\tau)$ is identical to that given in equation (S14) and (S15). The likelihood of the tree is given by the sum of the probability $D_{e,a}$ across states a weighted by the prior probility of those states.

$$\mathcal{L}(\mathcal{T}, \mathcal{C}_y | \Theta_{MBDS}) = \sum_a \pi_a D_{stem,a}(T) \tag{S25}$$

Stochastic mapping and integration over colourings

The likelihood of a given coloured tree is given by equation (S19). To calculate the likelihood of the tree where only the tip states are known as in the BiSSEness likelihood above, we must integrate over all possible colourings of the tree as specified in equation (S20). Such an integral can be performed by Monte Carlo integration, first simulating tree colorings consistent with the observed tip states from the distribution $Pr(\mathcal{C}|\mathcal{C}_y\Theta_{MBDS})$. By then taking the mean of the likelihoods of each of these coloured trees we can approximate the tree likleihood with known tip states $\mathcal{L}\left(\mathcal{T}, \mathcal{C}_y|\Theta_{MBDS}\right)$. To sample such colourings we using a stochastic extend the stochastic mapping method first developed by (Nielsen, 2002) allowing for time-varying

transition rates and cladogenic state change.

Preliminaries

- Let $Q(t)$ be the transition rate matrix at time t such that $Q_{j,i}(t)$ is the rate at which lineages in state i (column) transitions to state j (row). For the multi-type diversification process this transition rate is given by:

$$Q_{j,i}(t) = \begin{cases} -\sum_{k=1}^{A} \left(\gamma_{i,k}(T-t) + \lambda_{i,k}(T-t)E_i(T-t)\right) & if\, i = j \\ \gamma_{i,j}(T-t) + \lambda_{i,j}(T-t)E_i(T-t) & if\, i \neq j \end{cases} \tag{S26}$$

  Note that the matrix Q is defined forward in time from the origin to the present whereas the diversification rates are defined backward in time. To allow for the computation of the transition probability matrix P defined below, we approximate the time-variable transition rates with a piecewise constant function such that for the interval $(l-1) * \Delta t < t < l\Delta t$ the transition rate matrix is assumed to be constant as given by:

$$Q_{j,i,l} = Q_{j,i}(l * \Delta t) \tag{S27}$$

  where $\Delta t = \frac{T}{n_L}$ is a small uniform time and $l \in \{1, 2, ...n_L\}$.

- The transition probability $P_{j,i,l}$ of transitioning from state i to state j between $(l-1) * \Delta$ and Δt can be approximated as:

$$P_{i,j,l} \approx \mathbf{I} + Q_{j,i,l}\Delta t + \mathcal{O}(\Delta t^2), \tag{S28}$$

  where $\mathbf{I}$ is the identity matrix.

- Let $P_{j,i}(t_0, t_F)$ be the probability of transitioning from state i at time t_0 to state j at time t_F. Define $L(t)$ as the integer index such that $(L(t)-1)\Delta t < t < L(t)\Delta t$.

$$\mathbf{P}(t_0, t_F) \approx \mathbf{P}_{L(t_f)} \cdot \mathbf{P}_{L(t_f)-1} \cdot ... \cdot \mathbf{P}_{L(t_0)} \tag{S29}$$

  Assuming Δt is small this approximation assumes that no events occur between times t_0 and $L(t_0)\Delta t$ and between t_F and $L(t_F)\Delta t$.

- Let D_K be the tip states of the descendants of node K.

- Let Y_K bet the state of node K.

- With respect to stochastic mapping an edge refers to a *topological edge*, spanning between two nodes in the tree topology $\mathcal{T}$. This is in contrast to an edge in the coloured tree, as used in equation (S16), which is a portion of topological edge that is all one colour/state.

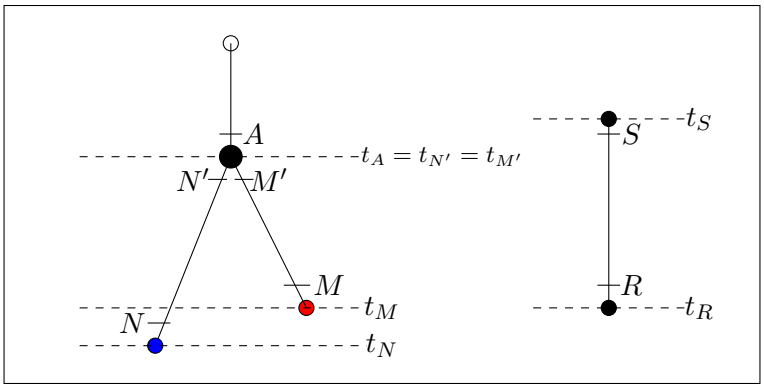

Figure S2: The two edge motifs in the tree with respect to stochastic mapping: A) a branching event where an edge A gives rise to a edge N and an edge M and B) the stem edge between the *stem node* at the tree origin S and the *root node* at the most recent common ancestor R.

Step 1: Calculate $f_{K,i} = Pr(D_K|Y_K = i)$

Define $f_{K,i} = Pr(D_K|Y_K = i)$, the probability of observing the known descendants of node K, D_K, given that node K is in state i.

For each tip node T:

$$f_{T,i} = \begin{cases} 1 & if\, Y_T = i \\ 0 & otherwise \end{cases} \tag{S30}$$

For each "originating" node (e.g, N' and M' in Figure S2):

$$\begin{aligned} f_{N',i} &= \sum_l P_{i,l}(t_{N'}, t_N) f_{N,l} \\ f_{M',i} &= \sum_l P_{i,l}(t_{M'}, t_M) f_{M,l} \end{aligned} \tag{S31}$$

For each "ancestral"/"ending" node (e.g., A in Figure S2)

$$f_{A,i} = \sum_{j,k} \tilde{\lambda}_{i,\{j,k\}}(t_A) f_{N',j} f_{M',k} \tag{S32}$$

where $\tilde{\lambda}_{i,\{j,k\}}$ is the probability that a lineage of type i gives rise to a linage of type j and a linage of type k given that a birth event occurs.

$$\begin{aligned} \tilde{\lambda}_{i,\{j,k\}}(t) &= \frac{\lambda_{i,\{j,k\}}(t)}{\sum_{a,b,c} \lambda_{a,\{b,c\}}(t)} \\ \lambda_{i,\{j,k\}} &= \begin{cases} \lambda_{i,i}(t) & if\, j = i \,\&\, k = i \\ \frac{\lambda_{i,i'}(t)}{2} & if\, j = i \,\&\, k = i' \\ \frac{\lambda_{i,i'}(t)}{2} & if\, j = i' \,\&\, k = i \\ 0 & if\, j = i' \,\&\, k = i' \end{cases} \end{aligned}$$

Using Felsenstein's pruning algorithm, equations (S30), (S31), and (S32) can be used to calculate the probability $f_{K,i}$ for each node K and each state i in the tree.

Step 2: Calculate $Pr(Y_K = j|D_K, Y_{K+1} = i)$+ simulate nodes, births

Let node $K+1$ be the direct ancestor of node K.

For the "stem" node:

$$Pr(Y_S = j|D) = \frac{f_{S,j}\pi_j}{\sum_i f_{S,i}\pi_i},$$

where $\pi_j = Pr(Y_S = j)$ is the probability the stem node has state j, hence this result follows from Bayes theorem.
Choose stem node state!.

For each "ending" node:

$$Pr(Y_N = j|D_N, Y_{N'} = i) = \frac{f_{N,j}P_{i,j}(t_{N'}, t_N)}{\sum_k f_{N,k}P_{i,k}(t_{N'}, t_N)}$$

The $f_{N,*}$ terms account for $Pr(Y_N = j|D_N)$ whereas the $P_{i,*}$ terms give the $Pr(Y_N = j|Y_{N'} = i)$.
Choose ending node state!.

For birth event and 'originating' node: Calculate the probability of the different birth events

$$Pr(Z_{A,j,k}|D_N, D_M, Y_A = i) = \frac{\lambda_{i,j,k}f_{N',j}f_{M',k}}{\sum_{b,c}\lambda_{i,b,c}f_{N',b}f_{M',c}}$$

Choose birth event!. Gives value of j and k.

Step 3: Simulate the colourings of each edge

Given the originating state of each node we can then us a Gillespie algorithm with the transition matrix P to simulate the colouring along each branch. Edge colourings are simulated until one is generated that matches the simulated ending state chosen in step 2. From the simulated edges and nodes we can obtain the vectors $\vec{x}_{a,b}$ and $\vec{w}_{a,b}$ as well as the initial state of the tree c^* and the states of the tip nodes as summarized by N_a.

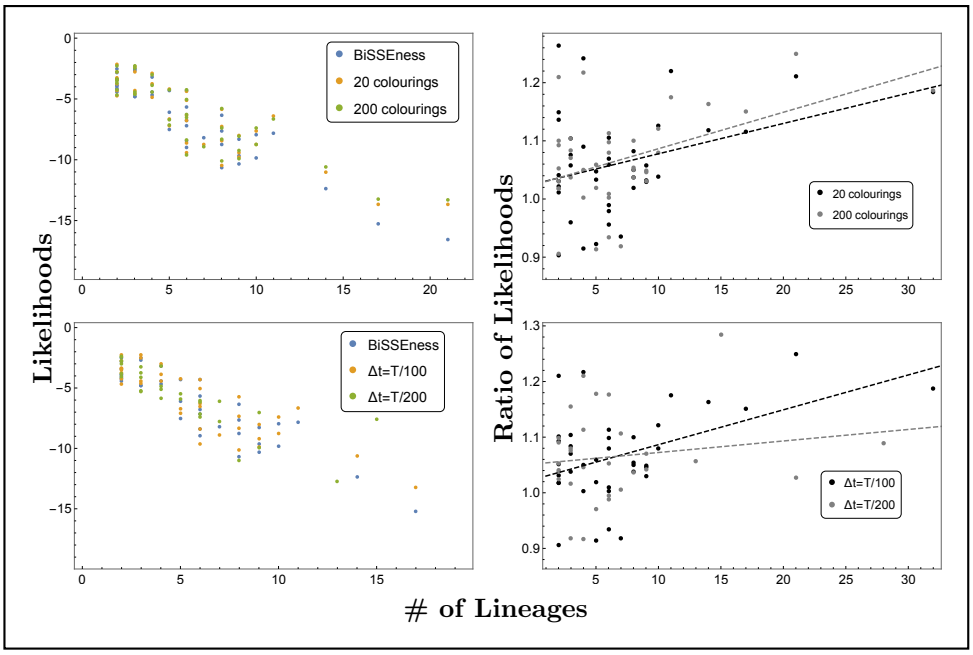

Figure S3: **Comparing likelihoods.** Comparing the tree likelihood calculated via Monte Carlo integration (see equation (S20)) (orange and green points) to likelihood calculated via direct numerical integration (equation (S25)). The error given by the ratio of the likelihoods as shown by the ratio of likelihoods in shown in the right hand plots. The accuracy of the likelihood is a foremost a function of tree size. The likelihood is reasonably accurate when relatively few terms are used in the Mote Carlo integration, e.g., averaging over only 20 colourings, and is improved by increasing the accuracy of the piecewise-constant approximation to the diversification rates (e.g., 200 vs. 100 piecewise-constant steps). Top row: $\Delta t = T/100$, bottom row: 200 colourings.

S4 Tables

Variable	Definition
A	The number of discrete lineage types.
$\lambda_{a,\{b,c\}}(\tau)$	The time-varying rate at which lineage of type a gives rise to daughter lineages of type b and c at time τ.
$\mu_a(\tau)$	The time varying rate at which lineages of type a go extinct at time τ
$\psi_a(\tau)$	The time varying rate at which lineages of type a are sampled.
$r_a(\tau)$	The time varying probability that upon sampling lineages are removed.
$\gamma_{a,b}(\tau)$	The time varying rate at which lineages of type a transition to type b.
ρ_a	The probability a lineage of type a alive at the present day is sampled.
N_a	The number of lineages of type a sampled at the present day.
$\vec{x}_{a,b,c}$	A vector of length $I_{a,b}$ giving the timing of speciation events where a lineage of type a gives rise to a daughter of type b.
$\vec{y}_a$	A vector of length J_a giving the times at which lineages of type a are sampled for which there are no sampled descendants (sampled tips).
$\vec{z}_a$	A vector of length K_a giving the times at which lineages of type a are sampled for which there *are* sampled descendants (sampled ancestors).
$\vec{w}_{a,b}$	A vector of length $L_{a,b}$ giving the times at which lineages of type a mutate to lineages of type b.
c	The state at the origin (stem node) of the phylogeny.
$\mathcal{T}$	The tree topology.
$\mathcal{C}$	The colouring of the tree.
$g_e(a,\tau)$	The probability an edge e of type a alive at time τ gives rise to the subsequently observed phylogeny
$E_a(\tau)$	The probability a lineage of type a alive at time τ has no sampled descendants between time τ and the present day.
$\iota_{a,b}$	An indicator variable with value 2 if $a = b$ and 1 otherwise.
$\mathcal{C}_y$	The known tip states
$\mathcal{C}_z$	The known states of sampled ancestors
Θ_{MBDS}	The parameters of the multi-type BDS model
SSE Algorithm	
π_c	The probability the stem node has state c
$D_{N,a}(\tau)$	The probability a "topological" edge N of type a at time τ gives rise to the subsequently observed phylogeny.

Table S4: **Notation for the general multi-type model**.Throughout t represents time moving forward from the origin ($t = 0$) of the tree to the tips ($t = T$) and τ represents time moving backward from the tips ($\tau = 0$) to the origin ($\tau = T$). The indices a, b denote lineage types and can take on integer values between 1 and A.