Bayesian phylodynamic inference of multi-type population trajectories using genomic data

Timothy G. Vaughan[1,2,*] and Tanja Stadler[1,2]
[1] Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland
[2] Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland
***Corresponding author:** timothy.vaughan@bsse.ethz.ch

Abstract

Phylodynamic methods provide a coherent framework for the inference of population parameters directly from genetic data. They are an important tool for understanding both the spread of epidemics as well as long-term macroevolutionary trends in speciation and extinction. In particular, phylodynamic methods based on multi-type birth-death models have been used to infer the evolution of discrete traits, the movement of individuals or pathogens between geographic locations or host types, and the transition of infected individuals between disease stages. In these models, population heterogeneity is treated by assigning individuals to different discrete types. Typically, methods which allow inference of parameters under multi-type birth-death models integrate over the possible birth-death trajectories (i.e. the type-specific population size functions) to reduce the computational demands of the inference. As a result, it has not been possible to use these methods to directly infer the dynamics of trait-specific population sizes, infected host counts or other such demographic quantities. In this paper we present a method which infers these multi-type trajectories with almost no additional computational cost beyond that of existing methods. We demonstrate the practicality of our approach by applying it to a previously-published set of MERS-CoV genomes, inferring the numbers of human and camel cases through time, together with the number and timing of spillovers from the camel reservoir. This application highlights the multi-type population trajectory's ability to elucidate properties of the population which are not directly ancestral to its sampled members.

Key words: phylodynamics, particle filter, epidemiology, Bayesian phylogenetics

Introduction

The strong connections between genetic diversity and population dynamics play an integral role in our ability to use modern sequencing technology to draw conclusions about populations at every scale. The field of phylodynamics offers a growing number of model-based approaches to inferring population dynamics from sequence data by way of phylogenies. While the term "phylodynamics" was originally conceived in the context of molecular epidemiology (Grenfell *et al.*, 2004), the techniques which are now described as phylodynamic methods are applied in a diverse range of contexts from paleontology (Gavryushkina *et al.*, 2017) to developmental biology (Stadler *et al.*, 2021).

One widely-used family of phylodynamic models are the so-called birth-death models (Kendall, 1948) under which the population dynamics are the result of a stochastic process parameterized by individual birth and death rates. Combined with an explicit sampling process (Stadler, 2010) and time-dependent rate shifts (Stadler *et al.*, 2012), these models enable genomic inference of the past dynamics of critical epidemiological and macroevolutionary parameters, such as transmission rates, speciation rates, and migration rates.

Beyond these parameters, the dynamics of the populations themselves are potentially of interest. In epidemics these correspond to incidence and prevalence dynamics, while for macroevolutionary applications of birth-death models they correspond to species richness trajectories. While earlier birth-death phylodynamic inference approaches traditionally integrated over these trajectories for computational efficiency reasons, recent approaches allow population trajectories to be jointly inferred with the birth-death parameters. The approach of Vaughan *et al.* (2019) employs a particle marginal Metropolis-Hastings algorithm to jointly sample parameters and population trajectories. This approach can be generalized to nonlinear birth-death models, but has the drawback that it can be extremely demanding computationally. The more recent approach of Manceau *et al.* (2021) is less flexible, but allows for direct and tractable computation of the posterior distribution of population size at a given time.

The above models and methods focus on neutrally-evolving, unstructured populations, where birth, death and sampling rates are the same for all co-existing individuals; assumptions which are inappropriate for many applications. This has spurred the development of a sub-family of birth-death phylodynamic models which allow population members to be associated with a discrete type whose two (Maddison *et al.*, 2007) or more (FitzJohn, 2012; Kühnert *et al.*, 2016) values may influence the birth, death and sampling rates to which they are subject. These models have been applied extensively to the study of

trait-dependent speciation and extinction (eg. Barido-Sottani *et al.*, 2020; FitzJohn, 2012; Maddison *et al.*, 2007; Rabosky, 2014) and to epidemiological questions, where the discrete types have been used to model the transmission dynamics of pathogens across locations (eg. Nadeau *et al.*, 2021), host types (eg. Grear *et al.*, 2017; Guinat *et al.*, 2022), and pathogen strains (eg. Loiseau *et al.*, 2023; Rasmussen and Stadler, 2019).

While the applications are diverse, existing multi-type inference methods (eg. FitzJohn, 2012; Kühnert *et al.*, 2016) only allow the inference of the type- (and potentially time-)dependent model parameters, in conjunction with the type dynamics of lineages directly ancestral to the sampled data. Unlike the neutral case, population trajectories cannot, to our knowledge, be readily inferred using existing multi-type methods—despite being present implicitly in the underlying type-dependent models. This is for the same reason that earlier neutral methods could not infer these trajectories: they are integrated out of the phylodynamic likelihood calculation for reasons of computational efficiency.

Here we present an approach to performing joint Bayesian inference of the phylogenetic tree, multi-type birth-death model parameters, and ancestral lineage types, together with type-specific population trajectories. Importantly, our approach allows this inference to be done *without noticeably increasing the computational complexity of the inference*; a feat which is accomplished by "mapping" trajectories onto combinations of trees and parameters already sampled using existing Markov chain Monte Carlo techniques. After introducing the approach and demonstrating the correctness of its implementation, we apply the method to a published set of MERS-CoV genomes curated by Dudas *et al.* (2018) belonging to the 2012–2015 MERS outbreak in the Arabian Peninsula. These authors used the results of a coalescent-based ancestral-state inference method (Vaughan *et al.*, 2014) to deduce the presence of camel to human spillover events in those lineages ancestral to the observed pathogen genomes. Our results demonstrate how the new approach enables Bayesian phylodynamic inference of host-specific prevalence dynamics, which yields information about the number and timing of camel to human spillover events in the broader population from which the samples were collected.

New Approaches

Our approach to inferring multi-type trajectories from sequence data is based on the combination of a multi-type extension to our particle filtering approach (Vaughan *et al.*, 2019) with a stochastic mapping approach (Freyman and Höhna, 2018) to the placement of ancestral type changes on a tree.

In this section, we describe the model assumptions, outline the new inference approach, and derive the necessary extensions to the particle filter.

Linear birth-death phylodynamic models

We focus here on a population composed of a discrete set of individuals, each belonging to one of a fixed number d of distinct types. This multi-type system is assumed to evolve according to a stochastic linear birth-death branching process with piecewise-constant rates. The process begins at time 0 with a single individual, whose type $i \in [1,d]$ is chosen randomly with probability f_i. While variations of this process have been described by numerous authors over the years, we here consider the specific form described by Kühnert *et al.* (2016) and by Scire *et al.* (2022).

We introduce the notation X_i to represent an individual of type i. Additionally, we use Y_i to represent a sample of type i. The individuals comprising the population are subject to the following named reactions:

Birth$_{ij}$: An individual of type i gives birth to a new individual of type j with rate λ_{ij}:

$$X_i \xrightarrow{\lambda_{ij}} X_i + X_j$$

Death$_i$: An individual of type i dies at rate μ_i:

$$X_i \xrightarrow{\mu_i} 0$$

Migration$_{ij}$: An individual of type i changes type to j at rate m_{ij}:

$$X_i \xrightarrow{m_{ij}} X_j$$

SampR$_{ij}$: An individual of type i is sampled and removed at rate $\psi_i r_i$, where ψ_i represents the overall sampling rate and r_i is the probability of removal on sampling:

$$X_i \xrightarrow{r_i \psi_i} Y_i$$

SampNR$_{ij}$: An individual of type i is sampled without removal at rate $\psi_i(1-r_i)$:

$$X_i \xrightarrow{(1-r_i)\psi_i} Y_i + X_i$$

The linearity of the process is reflected in the dependence of each of these reactions on only a single reagent. The rates associated with these reactions indicate the probability per unit time of individual X_i undergoing the corresponding reaction at time t. Each of the parameters $\lambda_{ij}, \mu_i, \ldots$ are assumed to be piecewise-constant functions of time.

In addition to these continuous reactions, the population is subject to synchronous sampling at the times $\vec{t}^c$. At a given synchronous sampling time t_v^c, every extant lineage of type i is sampled with probability ρ_i^v, and each sampled lineage is removed from the population with probability r_i^v. The resulting reactions

are given the parameterized name SyncSample($\vec{n}^{\mathrm{R}}$,$\vec{n}^{\mathrm{NR}}$), where n_i^{R} and n_i^{NR} are respectively the number of removed and non-removed samples of type i produced by the reaction.

A complete realization of the multi-type birth-death process would be a full, binary tree recording the ancestral relationships between all events and individuals that appeared in the system. However, we are only interested those properties of a given realization about which we can reasonably learn using genetic data associated with a subsample of the population. We thus consider only these elements: the population trajectory $\mathcal{E}$, the times and types of all samples $\mathcal{S}$, the phylogenetic tree $\mathcal{T}$ relating these typed and timed samples, and the phylogenetic tree whose edges are annotated with ancestral types, $\mathcal{T}_C$. (In the parlance of Kühnert *et al.* (2016), we refer to $\mathcal{T}$ as a *tip-typed tree* and $\mathcal{T}_C$ as an *edge-typed tree.*) In other words, we ignore the topology of those parts of the full tree not directly ancestral to the samples $\mathcal{S}$.

The population trajectory $\mathcal{E}$ is composed of the starting population state vector $\vec{N}_0$ together with a sequence of events ϵ_1,ϵ_2,... and a corresponding sequence of population states $\vec{N}_1$,$\vec{N}_2$,..., the latter describing the state of the population immediately following each event. An event ϵ consists of a time, t_ϵ, and the corresponding reaction, r_ϵ. (Synchronous sampling reactions always produce an event in the event sequence, even when no samples are generated.)

For what follows it will be helpful to define several vector-valued piecewise-constant auxiliary functions of time which can be uniquely derived from $\mathcal{E}$ and $\mathcal{T}_C$. First among these is $\vec{N}(t)$, whose elements $N_i(t)$ represent the size of the population associated with type i at time t. Secondly we define $\vec{S}(t)$, whose i^{th} element represents the cumulative number of samples collected from sub-population i up until time t. Finally, we define the function $\vec{K}(t)$, whose elements represent the number of extant lineages of $\mathcal{T}_C$ associated with each type at time t. These auxiliary functions, together with corresponding tip-typed $\mathcal{T}$ and edge-typed $\mathcal{T}_C$ trees, are illustrated in Figure 1.

Bayesian inference strategy

Our focus here is on jointly inferring population dynamics $\mathcal{E}$ and the edge-typed tree $\mathcal{T}_C$ from the number, timing and types of samples $\mathcal{S}$ produced by the population together with a genetic sequence alignment $\mathcal{A}$; the latter which is assumed to be the result of an unknown starting sequence evolving down the reconstructed tree $\mathcal{T}$ according to a standard Markovian nucleotide substitution model. That is, we seek to characterize the distribution $P(\mathcal{E},\mathcal{T}_C,\sigma,\eta|\mathcal{A},\mathcal{S})$, where σ and η are the parameters of the nucleotide substitution and birth-death models, respectively.

One approach to this problem would be to use Markov chain Monte Carlo (MCMC) directly to jointly sample the edge-typed tree $\mathcal{T}_C$, the multi-type population trajectory E, and all model parameters.

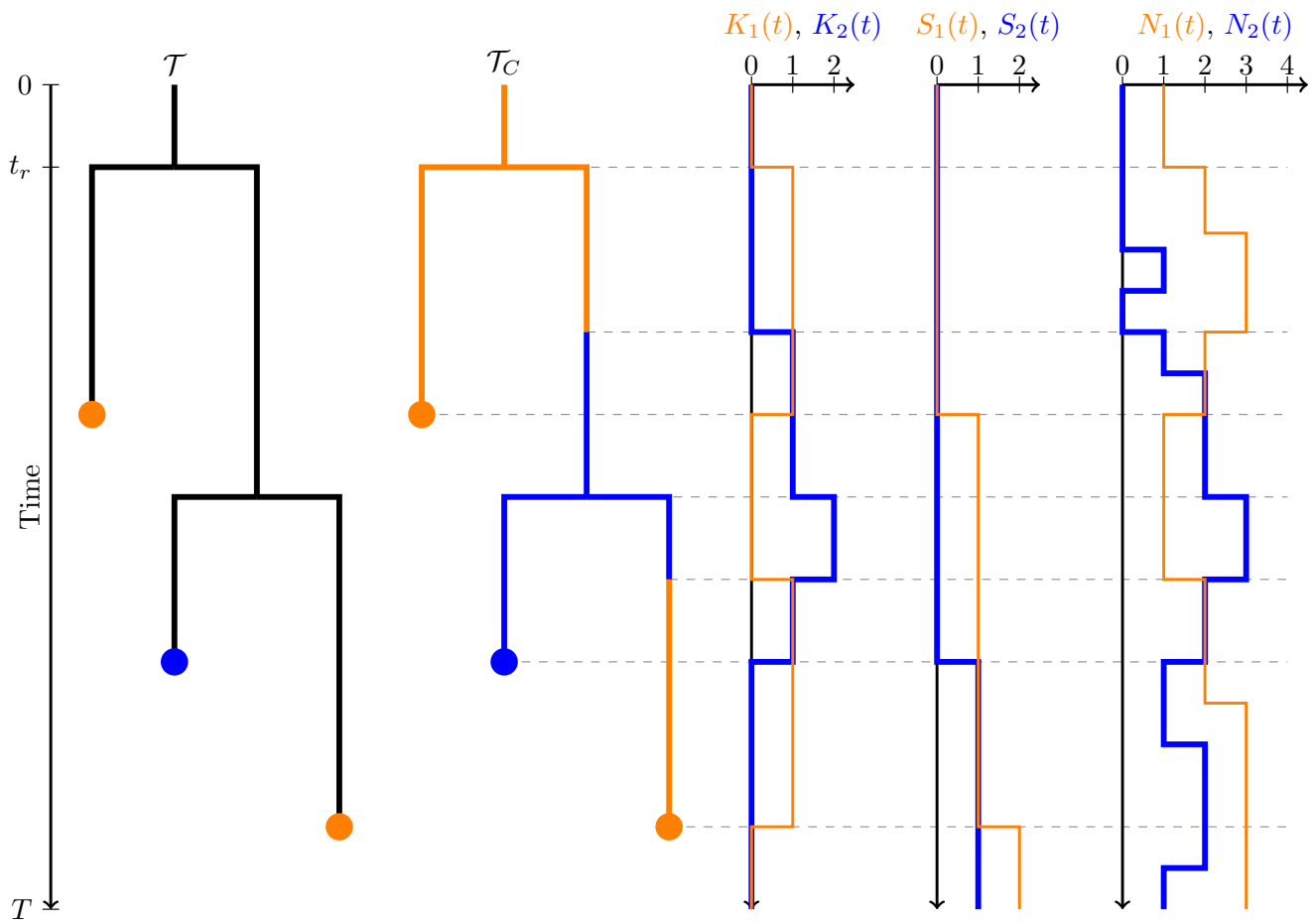

FIG. 1. Relationship between the tip-typed tree $\mathcal{T}$, the edge-typed tree $\mathcal{T}_C$, the multi-type lineage-through-time function $\vec{K}(t)$, the cumulative sample count function $\vec{S}(t)$ and the multi-type population size function $\vec{N}(t)$, for a single realization of a multi-type birth-death-sampling model starting at time 0 and ending at time T. The time of the root of the tree is t_r.

However, this approach would likely prove impractical due to the extremely large dimensionality of the state space. (While MCMC algorithms for directly sampling edge-typed trees under a variety of models are in use (eg. Beerli and Felsenstein, 2001; Ewing *et al.*, 2004; Kühnert *et al.*, 2016; Vaughan *et al.*, 2014), they are known to suffer from convergence issues which motivate the use of alternative schemes (De Maio *et al.*, 2015; Hey and Nielsen, 2007; Müller *et al.*, 2018). The addition of the highly-dimensional trajectory variable $\mathcal{E}$ to this space would likely make this problem worse.)

Instead, the approach that we advocate here is analogous to the general approach championed by Hey and Nielsen (2007) to improve the efficiency of inference of migration rates under the isolation-migration coalescent model, and applied more recently to the inference of ancestral types under the multi-type birth-death model by Freyman and Höhna (2018). This approach still involves MCMC, but to sample a marginalized posterior of lower dimensionality that includes neither ancestral types nor population trajectories. Instead, these latent variables are independently drawn from their conditional probability distributions given the MCMC-sampled tip-typed trees and birth-death parameters.

The starting point for our inference strategy is the posterior distribution targeted by the MCMC algorithm developed by Kühnert *et al.* (2016):

$$P(\mathcal{T},\eta,\sigma|\mathcal{A},\mathcal{S}) = \frac{P(\mathcal{A}|\mathcal{T},\sigma)P(\mathcal{T},\mathcal{S}|\eta)P(\eta,\sigma)}{P(\mathcal{A},\mathcal{S})} \tag{1}$$

Here σ and η are the parameters of the nucleotide substitution and birth-death models, respectively. The phylogenetic likelihood $P(\mathcal{A}|\mathcal{T},\sigma)$ is determined by the a particular combination of molecular substitution and clock model whose parameters we denote using σ. The phylodynamic likelihood $P(\mathcal{T}|\mathcal{S},\eta)$ for the multi-type birth-death model is computed by numerically integrating a set of ordinary differential equations. Finally, the prior $P(\sigma,\eta)$ expresses prior information on the model parameters. (The denominator is the marginal likelihood for the model, whose computation is unnecessary when using MCMC sampling.)

To infer the multi-type birth-death model trajectories, we incorporate the appropriate conditional distributions in the following way:

$$P(\mathcal{E},\mathcal{T}_C,\eta,\sigma|\mathcal{A},\mathcal{S}) = P(\mathcal{E}|\mathcal{T}_C,\eta)P(\mathcal{T}_C|\mathcal{T},\eta)P(\mathcal{T},\eta,\sigma|\mathcal{A},\mathcal{S}) \tag{2}$$

The sampling approach involves three steps:

1. sampling $\mathcal{T},\eta,\sigma \sim P(\mathcal{T},\eta,\sigma|A,\mathcal{S})$ using MCMC as described by Kühnert *et al.* (2016),

2. sampling $\mathcal{T}_C \sim P(\mathcal{T}_C|\mathcal{T},\eta)$ using a stochastic mapping algorithm for ancestral state reconstruction analogous to the approach developed by Freyman and Höhna (2018), and

3. sampling $\mathcal{E} \sim P(\mathcal{E}|\mathcal{T}_C,\eta)$ using a particle filtering algorithm.

The trajectories sampled in this way characterise the marginal distribution $P(\mathcal{E}|\mathcal{A},\mathcal{S})$.

A key advantage of this approach is that the second and third steps need only be applied to a small subset of the samples of $\mathcal{T},\eta,\sigma$ produced by the MCMC chain. This is because the sequence of states produced by MCMC algorithms are usually highly autocorrelated, meaning that a small subset of the states visited can characterize the target distribution as well as the full chain. While there is no well-defined consensus on the number of "effectively independent" samples required to produce useful characterizations of phylodynamic posteriors (although for parameter estimates it is common to treat values above 200 as acceptable), it is clear that this number is usually many orders of magnitude smaller than the number of iterations needed by the MCMC algorithm to achieve it. Thus, the ability to apply steps 2 and 3 to this subset means that these components of the approach are almost cost-free relative to the far larger computational cost of the MCMC portion of the algorithm.

As stated above, the MCMC algorithm we use for step 1 above is precisely that presented by Kühnert *et al.* (2016), thus we do not describe it further here. The following two sections describe the algorithms used to sample from the conditional distributions required for steps 2 and 3.

Stochastic mapping of ancestral types

Here we describe the approach we take to generating samples from $P(\mathcal{T}_C|\mathcal{T},\eta)$. While this approach is in many ways equivalent to the approach described by Freyman and Höhna (2018), we present here a more streamlined derivation and algorithm.

Consider a particular time t on a particular edge e of the tree $\mathcal{T}$. We use this point (e,t) to divide the tree into two components: $\mathcal{T}_{\downarrow}(e,t)$ being the clade descending from this point, and $\mathcal{T}_{\uparrow}(e,t)$ being the portion of the tree remaining when this clade is removed. By definition, these components are perfectly complementary in the sense that $\mathcal{T}=\mathcal{T}_{\downarrow}(t,e)\cup\mathcal{T}_{\uparrow}(e,t)$.

Defining $V(e,t)$ to be the (unknown) ancestral type value at (e,t), we can express the marginal probability that this trait takes a particular value i as

$$P(V(e,t)=i|\mathcal{T},\eta)=\frac{P(\mathcal{T}_{\uparrow}(e,t),V(e,t)=i|\eta)P(\mathcal{T}_{\downarrow}(e,t)|V(e,t)=i,\eta)}{P(\mathcal{T}|\eta)} \tag{3}$$

Noting that $P(\mathcal{T}_{\downarrow}(e,t)|V(e,t)=i,\eta)=g_i^e(t)$ where $g_i^e(t)$ is the partial tree probability function considered by Kühnert *et al.* (2016), and using the notation $h_i^e(t)=P(\mathcal{T}_{\uparrow}(e,t),V(e,t)=i|\eta)$ and $\pi_i^e(t)=P(V(e,t)=i|\mathcal{T},\eta)$, we can write this as

$$\pi_i^e(t)=\frac{1}{Z}h_i^e(t)g_i^e(t), \tag{4}$$

where $Z=P(\mathcal{T}|\eta)$ is the normalization constant.

Our goal is to determine a stochastic process which produces realizations of $V(e,t)$ for every point in the tree. To do this, we use the above expansion to derive the master equation for the marginal ancestral type distribution by directly differentiating this equation:

$$\dot{\pi}_i^e(t)=\frac{1}{Z}(\dot{h}_i^e(t)g_i^e(t)+h_i^e(t)\dot{g}_i^e(t)) \tag{5}$$

As described by Kühnert *et al.* (2016) and others, the time derivative for the partial tree probability $g_i^e(t)$ along an edge is given by the following system of ordinary differential equations (ODEs):

$$\dot{g}_i^e(t)=(\sum_j(\lambda_{ij}+m_{ij})+\mu_i+\psi_i)g_i^e(t)-\sum_j\lambda_{ij}(g_i^e(t)p_j(t)+g_j^e(t)p_i(t))-\sum_j m_{ij}g_i^e(t), \tag{6}$$

where $p_i(t)$ is the probability that an individual of type i alive at time t yields no sampled descendants, which follows its own ODE also described by Kühnert *et al.* (2016). (The synchronous sampling

probabilities appear only in the boundary conditions of these ODEs, as they do not affect the continuous-time evolution of the tree probabilities.)

We now make use of the normalization constraint that $\frac{d}{dt}\sum_i \pi_i^e(t) = \sum_i \dot{\pi}_i^e(t) = 0$ to give

$$\sum_i \dot{h}_i^e(t) g_i^e(t) = -\sum_i h_i^e(t) \dot{g}_i^e(t) \tag{7}$$

Substituting equation 6 into the right-hand side and performing a change of variables on the summation index, this becomes

$$\sum_i \dot{h}_i^e(t) g_i^e(t) = \sum_i \left[-\left(\sum_j (\lambda_{ij} + m_{ij}) + \mu_i + \psi_i \right) h_i^e(t) + \sum_j (\lambda_{ij} h_i^e(t) p_j(t) + \lambda_{ji} p_j(t) h_j^e(t) + m_{ji} h_j^e(t)) \right] g_i^e(t). \tag{8}$$

For this equation to hold for arbitrary $g_i^e(t)$, the summation kernels must be equal. Thus

$$\dot{h}_i^e(t) g_i^e(t) = \left[-\left(\sum_j (\lambda_{ij} + m_{ij}) + \mu_i + \psi_i \right) h_i^e(t) + \sum_j (\lambda_{ij} h_i^e(t) p_j(t) + \lambda_{ji} p_j(t) h_j^e(t) + m_{ji} h_j^e(t)) \right] g_i^e(t). \tag{9}$$

Substituting this, together with equation 6 back into equation 5 yields

$$\begin{aligned} \dot{\pi}_i^e(t) &= \frac{1}{Z} \sum_j \left(\lambda_{ji} p_j(t) h_j^e(t) g_i^e(t) - \lambda_{ij} p_i(t) h_i^e(t) g_j^e(t) + m_{ji} h_j^e(t) g_i^e(t) - m_{ij} h_i^e(t) g_j^e(t) \right) \\ &= \sum_j \left[Q_{ji}^e(t) \pi_j^e(t) - Q_{ij}^e(t) \pi_i^e(t) \right] \end{aligned} \tag{10}$$

where we have defined

$$Q_{ij}^e(t) = (\lambda_{ij} p_i(t) + m_{ij}) \frac{g_j^e(t)}{g_i^e(t)}. \tag{11}$$

We thus find that, conditional on the tree and parameter values, the ancestral type evolves down the tree according to a continuous-time Markov process with time-dependent transition rates $Q_{ij}^e(t)$. Furthermore, using simple arguments (see Supplemental Text) we can identify the transition probability for changes at internal state nodes, $W_{ijk}^e = \tilde{W}_{ijk}^e / \sum_{j'k'} \tilde{W}_{ij'k'}^e$, where

$$\tilde{W}_{ijk}^e = g_j^{e_l}(t_e) g_k^{e_r}(t_e) \begin{cases} \frac{1}{2}\lambda_{ik} & j = i, k \neq i \\ \frac{1}{2}\lambda_{ij} & j \neq i, k = i \\ \lambda_{ii} & j = i, k = i \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

(Such state changes are possible when the off-diagonal elements of the birth rate matrix λ_{ij} are non-zero.)

Since we can pre-compute the components of these rates and transition probabilities by numerically solving the differential equations given by Kühnert *et al.* (2016), we can directly sample the state changes by (a) sampling the starting state from its marginal posterior, then (b) simulating this Markov process forwards in time down the tree.

Sampling the appropriately conditioned starting state can be done directly by noting that at the start of the root edge r, $h_i^r(0) = P(\mathcal{T}_\uparrow(r,0), V(r,0) = i|\eta)$ reduces to $P(V(r,0) = i|\eta) = f_i$, and thus that we can compute the marginal probabilities for each possible starting state $\pi_i^r(0) = f_i g_i^r(0) / \sum_{i'} f_{i'} g_{i'}^r(0)$ directly.

Simulating the process and hence drawing a sample from $P(\mathcal{T}_C|\mathcal{T},\eta)$ can be done using any algorithm capable of simulating trajectories from a continuous-time Markov chain with time-dependent rates. The approach we use, given as algorithm 1, is a simple extension of the original Stochastic simulation algorithm (Gillespie, 1976).

Importance sampling trajectories

While it is straightforward to produce samples of the multi-type trajectories from the unconditioned distribution $P(\mathcal{E}|\eta)$ through direct stochastic simulation (Gillespie, 1976), our goal is to sample $\mathcal{E}$ from the conditional distribution $P(\mathcal{E}|\mathcal{T}_C,\eta)$. We will bridge this gap using sequential importance sampling.

A naïve importance sampler might use direct simulation to sample an ensemble of $M \gg 1$ trajectories directly from $P(\mathcal{E}|\eta)$ and weight each trajectory using the edge-typed tree probability $P(\mathcal{T}_C|\mathcal{E})$. However, literal application of this method is not practical, since $P(\mathcal{T}_C|\mathcal{E}^{(m)})$ will be zero for any $\mathcal{E}^{(m)} \sim P(\mathcal{E}|\eta)$. This is because the tree density is only non-zero when the real-valued times of all events represented on the tree coincide precisely with the times of events in $\mathcal{E}^{(m)}$.

For this reason, we instead simulate trajectories from the modified distribution $P^*(\mathcal{E}|\eta)$. This distribution corresponds to a similar process as $P(\mathcal{E}|\eta)$ but with the important distinction that all events contained in the edge-typed tree $\mathcal{T}_C$ occur with probability 1, and any events that are incompatible with $\mathcal{T}_C$ are forbidden. Weighting these trajectories with the unnormalized weight function

$$w(\mathcal{E}) = P(\mathcal{T}_C|\mathcal{E}) \frac{P(\mathcal{E}|\eta)}{P^*(\mathcal{E}|\eta)} \tag{13}$$

produces our desired target distribution, which can be seen by noting that:

$$\frac{w(\mathcal{E}) P^*(\mathcal{E}|\eta)}{\int w(\mathcal{E}') P^*(\mathcal{E}') d\mathcal{E}'} = \frac{P(\mathcal{T}_C|\mathcal{E}) P(\mathcal{E})}{\int P(\mathcal{T}_C|\mathcal{E}') P(\mathcal{E}'|\eta) d\mathcal{E}'} = P(\mathcal{E}|\mathcal{T}_C,\eta).$$

To compute these weights, we need to specify three distributions: the trajectory probability $P(\mathcal{E}|\eta)$, the importance trajectory distribution $P^*(\mathcal{E}|\eta)$ and the tree probability $P(\mathcal{T}_C|\mathcal{E})$.

The trajectory distribution follows directly from the definition of the birth-death process, and can be written

$$P(\mathcal{E}|\eta) = \exp\left[-\int_0^T \left(\sum_{r \in R} A_r(t) \right) dt \right] \prod_{\epsilon \in \mathcal{E}} A_{r_\epsilon}(t_\epsilon) \tag{14}$$

where R represents the set of all possible continuous reactions in the model (i.e. excluding synchronous sampling reactions). For such reactions, $A_r(t)$ is the piecewise-constant propensity of reaction r at time

Algorithm 1: Stochastic mapping algorithm. The notation ChildCount(e) represents the number of child edges below edge e; Child(e) is the single child edge when ChildCount(e)=1; and LeftChild(e) and RightChild(e) are the left and right child edges when ChildCount(e)=2.

Input:

$g^e(t)$: the partial tree probabilities computed as in Kühnert *et al.* (2016)

$p_i(t)$: the clade non-sampling probabilities computed as in Kühnert *et al.* (2016)

T: the total process time

η: the piecewise-constant birth-death model parameters

Initialize:

Sample s_0 from $f_{s_o} g^r_{s_0}(0) / \sum_j f_j g^r_j(0)$ for root edge r

Call `Map(`r,0,s_0`)`

Function `Map(`e,t,s`)`**:**

```
while t < t_e do
    Sample u from Unif(0,1).
    Numerically solve exp[ - ∫_t^{t'} Σ_i Q^e_{si}(t'') dt'' ] = u for t'.

    if t' > t_e then
        Set t ← t_e.
        switch ChildCount(e) do
            case 0 do return
            case 1 do Call Map(Child(e),t_e,s)
            case 2 do
                Sample (s_l,s_r) from W^e_{s,s_l,s_r}.
                Call Map(LeftChild(e),t_e,s_l)
                Call Map(RightChild(e),t_e,s_r)
        end
    else
        Sample s' from Q^e_{ss'} / Σ_{s''} Q^e_{ss''}.
        Set t ← t' and s ← s'.
    end
end
```

t (e.g. for $A_{\mathrm{Birth}_{ii}}(t) = \lambda_{ii} N_i(t)$). For synchronous sampling reactions on the other hand, $A_r(t^c_v)$ is the multinomial probability of the particular removed/non-removed synchronous sampling outcome at the

sampling time t_v^c, that is:

$$A_{\mathrm{SyncSamp}(\vec{n}^{\mathrm{R}},\vec{n}^{\mathrm{NR}})}(t_v^c)=\prod_i \binom{N_i(t_v^c)}{n_i^{\mathrm{R}},n_i^{\mathrm{NR}}}(1-\rho_i)^{N_i(t_v^c)-(n^{\mathrm{R}_i}+n^{\mathrm{NR}_i})}(r_i^v\rho_i^v)^{n_i^{\mathrm{R}}}((1-r_i^v)\rho_i^v)^{n_i^{\mathrm{NR}}}. \tag{15}$$

In both cases, r_ϵ is the reaction involved in event ϵ, and t_ϵ is the time of that event.

The probability of the edge-typed tree can be similarly decomposed:

$$P(\mathcal{T}_C|\mathcal{E})=C\prod_{\epsilon\in\mathcal{E}} f_{r_\epsilon}(t_\epsilon) \tag{16}$$

where C is a normalizing constant and $f_r(t)\equiv P_r(\vec{K}(t_\epsilon^-),\vec{S}(t^-)|\vec{K}(t^+),\vec{S}(t^+),\vec{N}(t^+))$, the probability for a reaction r to effect a specific instantaneous change in the lineages-through-time function and cumulative sample count function at time t. The times t^- and t^+ refer to the times immediately before and after t. The functions $\vec{K}(t)$ and $\vec{S}(t)$ are the lineages-through-time and cumulative sample count functions for $\mathcal{T}_C$, and $\vec{N}(t)$ is the population size function for $\mathcal{E}$—as illustrated in Figure 1.

The normalizing constant C accounts for the fact that, in general, many unique tree topologies may to a given pair of lineages-through-time and cumulative sample count functions. Since this number affects only the overall normalization of the weight function, it plays no role in our calculation and can be safely ignored.

The probabilities $f_r(t)$ are derived by considering the number of ways that lineages belonging to the tree can be involved in a given reaction, and are given below for each of the reactions in our model. (In the following, $\hat{i}$ and $\hat{j}$ represent unit vectors of dimension d where the ith or, respectively, j^{th} element is 1 and all other elements are 0.)

- Single-type birth reactions:

$$f_{\mathrm{Birth}_{ii}}(t)=\begin{cases}\binom{K_i(t^+)}{2}/\binom{N_i(t^+)}{2} & \vec{K}(t^-)=\vec{K}(t^+)-\hat{i},\vec{S}(t^-)=\vec{S}(t^+)\\ 1-\binom{K_i(t^+)}{2}/\binom{N_i(t^+)}{2} & \vec{K}(t^-)=\vec{K}(t^+),\vec{S}(t^-)=\vec{S}(t^+)\\ 0 & \text{otherwise}\end{cases} \tag{17}$$

- Between-type birth reactions:

$$f_{\mathrm{Birth}_{ij}}(t)=\begin{cases}\frac{K_i(t^+)K_j(t^+)}{N_i(t^+)N_j(t^+)} & \vec{K}(t^-)=\vec{K}(t^+)-\hat{j},\vec{S}(t^-)=\vec{S}(t^+)\\ \left(1-\frac{K_i(t^+)}{N_i(t^+)}\right)\frac{K_j(t^+)}{N_j(t^+)} & \vec{K}(t^-)=\vec{K}(t^+)-\hat{j}+\hat{i},\vec{S}(t^-)=\vec{S}(t^+)\\ 1-\frac{K_j(t^+)}{N_j(t^+)} & \vec{K}(t^-)=\vec{K}(t^+),\vec{S}(t^-)=\vec{S}(t^+)\\ 0 & \text{otherwise}\end{cases} \tag{18}$$

- Migration reactions:

$$f_{\text{Migration}_{ij}}(t)=\begin{cases} K_j(t^+)/N_j(t^+) & \vec{K}(t^-)=\vec{K}(t^+)-\hat{j}+\hat{i}, \vec{S}(t^-)=\vec{S}(t^+) \\ 1-K_j(t^+)/N_j(t^+) & \vec{K}(t^-)=\vec{K}(t^+), \vec{S}(t^-)=\vec{S}(t^+) \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

- Death reactions:

$$f_{\text{Death}_i}(t)=\begin{cases} 1 & \vec{K}(t^-)=\vec{K}(t^+), \vec{S}(t^-)=\vec{S}(t^+) \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

- Sampling reactions with removal:

$$f_{\text{SampR}_i}(t)=\begin{cases} 1 & \vec{K}(t^-)=\vec{K}(t^+)+\hat{i}, \vec{S}(t^-)=\vec{S}(t^+)-\hat{i} \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

- Sampling reactions without removal:

$$f_{\text{SampNR}_i}(t)=\begin{cases} K_i(t^+)/N_i(t^+) & \vec{K}(t^-)=\vec{K}(t^+), \vec{S}(t^-)=\vec{S}(t^+)-\hat{i} \\ 1-K_i(t^+)/N_i(t^+) & \vec{K}(t^-)=\vec{K}(t^+)+\hat{i}, \vec{S}(t^-)=\vec{S}(t^+)-\hat{i} \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

- The probability due to synchronous sampling events is a product of hypergeometric distributions for the number of sampled ancestors of each type produced by a synchronous sampling reaction at a given time:

$$f_{\text{SyncSamp}(\vec{n}^{\text{R}},\vec{n}^{\text{NR}})}(t)=\begin{cases} \prod_i \frac{\binom{K_i(t^+)}{n_i^{\text{SA}}(t)}\binom{N_i(t^+)-K_i(t^+)}{n_i^{\text{NR}}-n_i^{\text{SA}}(t)}}{\binom{N_i(t^+)}{n_i^{\text{NR}}}} & \vec{S}(t^+)-\vec{S}(t^-)=\vec{n}^{\text{NR}}+\vec{n}^{\text{R}}, 0\leq\vec{K}(t^-)-\vec{K}(t^+)\leq\vec{n}^{\text{NR}}+\vec{n}^{\text{R}} \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

  (Here $n_i^{\text{SA}}(t):=(S_i(t^+)-S_i(t^-))-(K_i(t^-)-K_i(t^+))$, i.e. the difference between the number of type-i samples generated at time t and the number of new type-i lineages appearing at the same time, represents the number of "sampled ancestors" of type i generated by the n_i^{NR} unremoved samples.)

Note that these tree event probabilities are a superset of those used by the particle filtering algorithm presented by Vaughan *et al.* (2019) and are equivalent to a special case of the general formalism presented by King *et al.* (2022, 2024).

In order to express the importance distribution, we partition the trajectory $\mathcal{E}=\mathcal{U}\cup\mathcal{O}$, where $\mathcal{O}$ includes only the events occurring at exactly the same time as events on the tree ("observed" events), and $\mathcal{U}$ contains the remaining ("unobserved") events. The importance distribution corresponds to a modified

birth-death process in which unobserved events occur at rate $A_r(t)f_r(t)$, while observed events occur with probability 1 and with a reaction r_ϵ chosen with probability $A_{r_\epsilon}(t_\epsilon)/\sum_{r'\in R(t_\epsilon)}A_{r'}(t_\epsilon)$ from the set $R(t_\epsilon)$ of reactions (including synchronous sampling reactions) compatible with the tree at the observed event time t_ϵ. This modified process ensures that (a) only those unobserved events compatible with the tree are included, and that (b) observed events occur with probability 1 and are associated with reactions also compatible with the tree. The probability for a trajectory generated by this process is written:

$$P^*(\mathcal{E}|\eta)=\exp\left[-\int_0^T\left(\sum_{r\in R}A_r(t)f_r(t)\right)dt\right]\prod_{\epsilon\in\mathcal{U}}A_{r_\epsilon}(t_\epsilon)f_{r_\epsilon}(t_\epsilon)\prod_{\epsilon\in\mathcal{O}}\frac{A_{r_\epsilon}(t_\epsilon)}{\sum_{r'\in R(t_\epsilon)}A_{r'}(t_\epsilon)} \tag{24}$$

Substituting this, together with the expressions for the tree and trajectory probabilities, into equation (13) yields the importance weight function:

$$w(\mathcal{E})=\exp\left[-\int_0^T\left(\sum_{r\in R}A_r(t)(1-f_r(t))\right)dt\right]\prod_{\epsilon\in\mathcal{O}}f_{r_\epsilon}(t_\epsilon)\sum_{r\in R(t_\epsilon)}A_r(t_\epsilon). \tag{25}$$

Thus, one could sample trajectories from $P(\mathcal{E}|\mathcal{T}_C,\eta)$ by sampling a large number of trajectories from $P^*(\mathcal{E}|\eta)$, computing their associated weights using $w(\mathcal{E})$, then sampling from this weighted distribution.

Particle filtering algorithm for multi-type trajectories

While the above strategy will produce the correct distribution in the limit of an infinitely large trajectory ensemble, in practice we are restricted to finite ensembles of trajectories. This means that samples drawn from the weighted ensemble will only approximately match the statistics of $P(\mathcal{E}|\mathcal{T}_C,\eta)$, to an extent governed by the effective sample size of the weighted ensemble. An intuitive way to think about this problem is to consider that the trajectories produced by the single-step importance sampler above are only assigned weights once they have been completely simulated, meaning that the ensemble is likely to include many trajectories with relatively low weights, each of which are no less expensive to simulate than highly-weighted trajectories.

One widely-used strategy for improving the effective sample size of the weighted ensemble is known as Sequential Monte Carlo or particle filtering. (Refer to Cappe *et al.* (2007) for a good overview of this approach.) This involves subdividing the single importance sampler above into a number of intermediate importance samplers, with resampling steps in between. The resampling helps keep disparities between trajectory weights in the ensemble manageable.

We apply this here by subdividing the trajectory $\mathcal{E}$ into portions $\mathcal{E}_1,\ldots,\mathcal{E}_L$, where each $\mathcal{E}_l$ includes only those trajectory events with times satisfying $t_\epsilon\in(t_{l-1},t_l]$ and where t_l correspond to the times of the $L=|\mathcal{O}|$ events observed on the tree. We similarly partition the set of unobserved events $\mathcal{U}$ into $\mathcal{U}_1,\ldots,\mathcal{U}_L$. Since (a) the stochastic process described by the importance distribution $P^*(\mathcal{E}|\eta)$ is Markovian, and (b)

the weight function $w(\mathcal{E})$ has a simple product form, we can write

$$P^*(\mathcal{E}_l|\mathcal{E}_{l-1},\eta) = \exp\left[-\int_{t_{l-1}}^{t_l}\left(\sum_{r\in R} A_r(t) f_r(t)\right)dt\right]\left[\prod_{\epsilon\in\mathcal{U}_l} A_{r_\epsilon}(t_\epsilon) f_{r_\epsilon}(t_\epsilon)\right]\frac{A_{r_l}(t_l)}{\sum_{r'\in R(t_l)} A_{r'}(t_l)} \tag{26}$$

and

$$w_l(\mathcal{E}_l) = \exp\left[-\int_{t_{l-1}}^{t_l}\left(\sum_{r\in R} A_r(t)(1-f_r(t))\right)dt\right] f_{r_l}(t_l)\sum_{r\in R(t_l)} A_r(t_l). \tag{27}$$

where we have used r_l to refer to the reaction corresponding to the observed event ending interval l. Our particle filter works by initializing an ensemble of trajectories (particles) based on the initial type defined by the multi-type tree $\mathcal{T}_c$ and then alternating between (1) propagating the ensemble to the next interval using stochastic simulation to sample from $P^*(\mathcal{E}_l|\mathcal{E}_{l-1},\eta)$ defined in Eq. (26), and (2) resampling the ensemble according to the weights $w_l(\mathcal{E}_l)$ defined in Eq. (27).

The full procedure is described in Alg. (2), and—in the large particle number limit—produces a single multi-type trajectory drawn from the true conditional distribution of trajectories given an edge-typed tree and corresponding multi-type birth-death parameters. This algorithm can be regarded as special case of Algorithm 3 in (Cappe *et al.*, 2007). However, as noted in our previous paper (Vaughan *et al.*, 2019) and discussed recently in more detail by King *et al.* (2022), when applied to such a genealogical process the weighted distribution of trajectories is only interpretable as a true conditional distribution given the tree once the entire tree has been considered.

Results

Implementation

The algorithms described were implemented as an extension to the Bayesian phylogenetic inference platform BEAST 2 (Bouckaert *et al.*, 2019). Specifically, these algorithms were implemented in the package BDMM-Prime, which is a recent revision of the original BDMM package published by Kühnert *et al.* (2016). Beyond stochastic mapping and multi-type trajectory inference, this package also makes use of efficiency and numerical stability improvements introduced by Scire *et al.* (2022) and the analytical single-type functionality of the earlier BDSKY (Stadler *et al.*, 2012) package, meaning that all analyses possible in these earlier packages are accessible within BDMM-Prime. It also includes an extensive and redesigned graphical interface, making it possible for users to configure more complex analyses compared with the earlier implementations. The new package is free software and is distributed under the GNU General Public License version 3. The source code is available from its GitHub repository (`https://github.com/tgvaughan/BDMM-Prime`), and a complete online manual including a tutorial demonstrating its use can be found at the project website (`https://tgvaughan.github.io/BDMM-Prime`).

Algorithm 2: Particle filtering algorithm used to sample a trajectory from $P(\mathcal{E}|\mathcal{T}_C,\eta)$.

Input:

$\mathcal{T}_C$: the edge-typed tree (output of the stochastic mapping algorithm)

η: the piecewise-constant birth-death model parameters

M: the number of particles used in the importance sampler

Initialize:

Define $\mathcal{E}_0^{(p)}$ to be the initial type of $\mathcal{T}_C$.

for $l \in [1,L]$ **do**

  for $p \in [1,M]$ **do**

    Sample $\tilde{\mathcal{E}}_l^{(p)} \sim P^*(\mathcal{E}_l|\mathcal{E}_{l-1}^{(p)},\eta)$ using either the Gillespie algorithm or τ-leaping.

    Compute $w_l^{(p)} = w_l(\tilde{\mathcal{E}}_l^{(p)})$.

  end

  for $p \in [1,M]$ **do**

    Compute normalized weight $\tilde{w}_l^{(p)} = w_l^{(p)} / \sum_{p' \in [1,M]} w_l^{(p')}$.

  end

  for $p \in [1,M]$ **do**

    Sample $p' \sim \text{Multinomial}(M;\tilde{w}^{(1)},...,\tilde{w}^{(M)})$.

    Define $\mathcal{E}_l^{(p)} = \tilde{\mathcal{E}}_l^{(p')}$ and $\alpha_l(p) = p'$.

  end

end

return $\mathcal{E} = \mathcal{E}_L^{(0)} \cup \mathcal{E}_{L-1}^{(\alpha_L(0))} \cup \mathcal{E}_{L-2}^{(\alpha_{L-1}(\alpha_L(0)))} \cup \mathcal{E}_{L-3}^{(\alpha_{L-2}(\alpha_{L-1}(\alpha_L(0))))} \ldots$.

Trait mapping validation tests

In order to validate the implementation of the stochastic mapping approach presented as algorithm 1, we simulated 10^5 edge-typed trees under a two-state model with $\lambda_{ij} = 2\delta_{ij} + (1-\delta_{ij})0.5$, $m_{ij} = 0.1$, $\mu_i = 1$, $\psi_i = 0.1$, $r_i = 0.5$ (for all i,j), and a simulation time of 5 time units. For each simulated edge-typed tree $\mathcal{T}_C$ we produced a corresponding tip-typed tree $\mathcal{T}$ by removing the ancestral type information. We then used the stochastic mapping algorithm to sample a new edge-typed tree $\mathcal{T}_{C'}$. If the stochastic mapping algorithm implementation correctly samples from $P(\mathcal{T}_C|\mathcal{T},\eta)$, the sample distributions of $\mathcal{T}_C$ and $\mathcal{T}_C'$ should be indistinguishable, since

$$\sum_{\mathcal{T}} P(\mathcal{T}_C|\mathcal{T},\eta) P(\mathcal{T}|\eta) = P(\mathcal{T}_C|\eta). \tag{28}$$

Figure 2 illustrates comparisons of distributions of summary two statistics computed from these samples: the number of type transitions along each tree, and the total edge lengths associated with each type. The

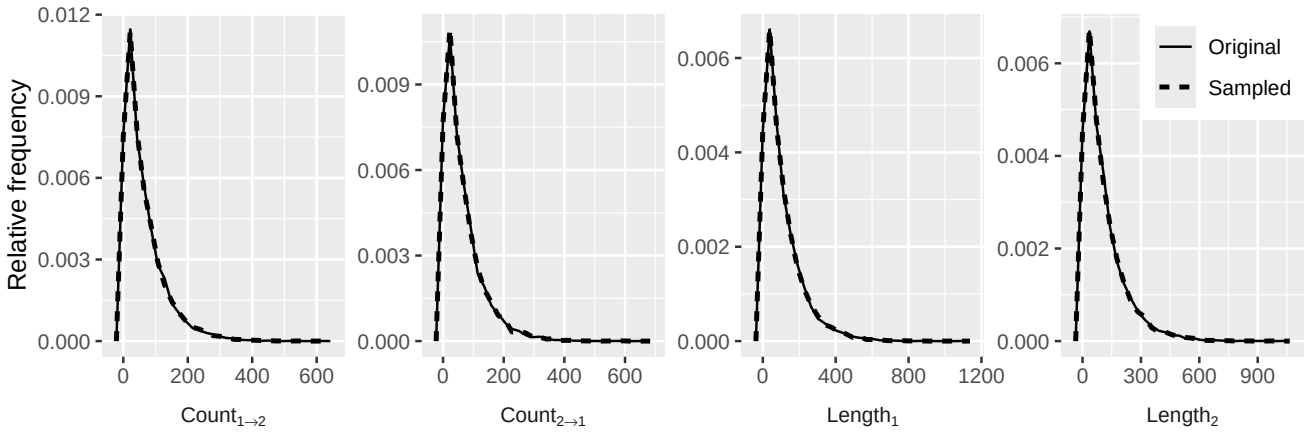

FIG. 2. Frequency polygons comparing the distribution of summary statistics for the edge-typed trees belonging to the original simulated distribution to the equivalent distribution computed from the edge-typed trees generated by the stochastic mapping algorithm. The statistic $\text{Count}_{i\to j}$ represents the number of transitions from type i to type j recorded by the edge-typed tree, while Length_i represents the total edge length associated with type i.

close agreement between these distributions provides strong evidence for the correctness of our algorithm and its implementation.

Trajectory sampling validation tests

We took a similar approach to validating the implementation of the trajectory particle filtering algorithm. We simulated an ensemble of 10^3 edge-typed trees $\mathcal{T}_C^{(m)}$ and trajectories $\mathcal{E}^{(m)}$ under the same two-type model used in the stochastic mapping validation above. For each tree in this ensemble we used the particle filter to sample $\mathcal{E}'^m \sim P(\mathcal{E}|\mathcal{T}_C^{(m)},\eta)$. Since

$$\sum_{\mathcal{T}_C} P(\mathcal{E}|\mathcal{T}_C,\eta)P(\mathcal{T}_C|\eta) = P(\mathcal{E}|\eta), \tag{29}$$

the two sample distributions $\mathcal{E}'^{(m)}$ and $\mathcal{E}^{(m)}$ should be identical, provided the particle filter used for the conditional sampling from $P(\mathcal{E}|\mathcal{T}_C,\eta)$ is correct.

Figure 3 shows a comparison between the these two trajectory distributions. Specifically, it illustrates the distribution of population sizes associated with each of the two types at each of a sequence of 51 evenly-spaced times between 0 and 5 (i.e. the start and end of the simulation window). The figure shows close agreement between the distribution of trajectories produced by direct simulation and the distribution produced by conditional sampling from the corresponding type-mapped trees. This agreement is strong evidence that the conditional multi-type trajectory sampling algorithm and its implementation are correct.

We also simulated a second ensemble of 10^3 edge-typed trees and multi-type trajectories under a similar but distinct model in which the continuous sampling was replaced with a pair of contemporaneous sampling events at times 4 and 5 following the start of the process. Details of this model and the

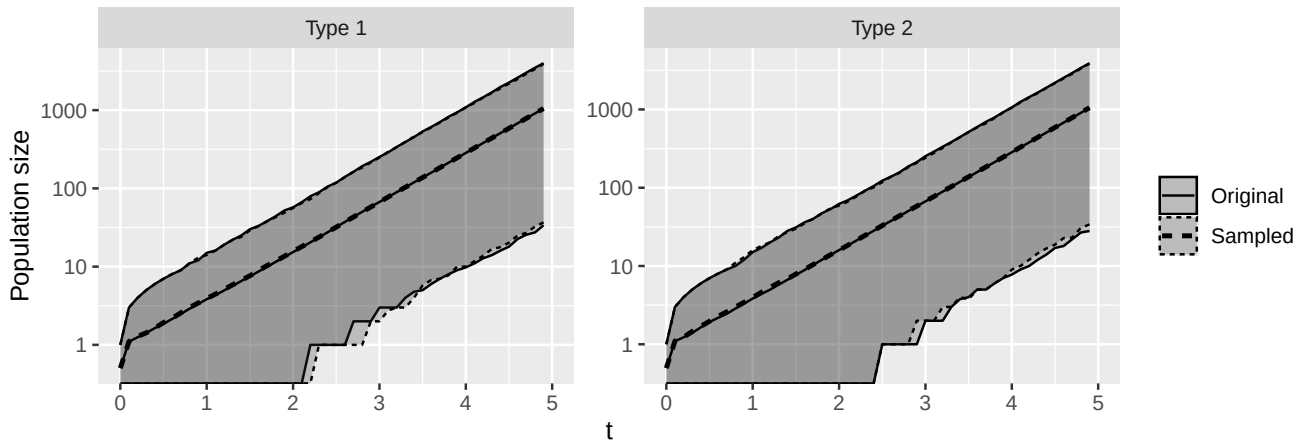

FIG. 3. Comparison of 2-type trajectory distributions produced by direct simulation of trajectories (solid lines) and trees with the distribution of 2-type trajectories sampled using the particle filter (dashed lines) to condition on the corresponding trees. The left and right panels compare the distributions of the type 1 and type 2 population sizes, respectively, each as a function of time. The central lines indicate the mean population size, while the thin lower and upper lines respectively indicate the 2.5% and 97.5% quantiles of the population size distribution.

comparison between the corresponding trajectory ensembles are shown in supplemental figure S1. This also shows agreement between the distributions, further supporting our claim of correct implementation.

Application: MERS in Camels and Humans

Following validation, we applied our approach to study the zoonotic transmission dynamics of the Middle East respiratory syndrome coronavirus (MERS-CoV). This virus was first isolated in 2012 (Zaki *et al.*, 2012) and has been responsible for localized outbreaks of a respiratory disease, primarily in or associated with the Arabian Peninsula, with a high rate of mortality (World Health Organization, 2022). While bats were initially believed to be the primary non-human reservoir, the pathogen was subsequently found to be endemic in camels in the region (see de Wit *et al.* (2016) for a comprehensive review).

We focused on the publicly-available MERS-CoV genomic dataset curated and used by Dudas *et al.* (2018) in their investigation of camel to human spillover dynamics. In that paper, the authors used a structured coalescent model (Vaughan *et al.*, 2014) to infer the timing and number of spillover events of the pathogen from camels to humans. Our goal here was to analyze same data using a multi-type birth-death model to jointly infer the two-host infection dynamics in the larger population, enabling direct quantification of the spillover dynamics.

The dataset is comprised of an alignment of 274 MERS-CoV genomes, including 100 genomes from camel hosts and 174 from humans. To avoid problems due to the recombination, we followed the approach taken by Dudas *et al.* (2018) in one of their analyses of considering only sites 21001–29364. We employed the nucleotide substitution model of Hasegawa *et al.* (1985) allowing for Gamma-distributed site to site rate heterogeneity (Yang, 1994). We assumed a relaxed clock model (Douglas *et al.*, 2021) with an informative prior on the mean rate based on the study of MERS-CoV evolution by Zhang *et al.* (2016).

Table 1. Parameter priors used in the analysis of the MERS-CoV data. Subscripts i and j specify host types selected from the possible values (Camel,Human). Note that d_s represents the duration in years between the first and last samples in the dataset.

Parameter	Unit	Prior distribution
R_{ii}	—	LogNorm(0,0.5)
R_{ij} $(i \neq j)$	—	Exp(1)
δ_i	Year^{-1}	LogNorm(3,0.5)
s_i	–	Unif(0,0.1)
r_i	–	1
T	Year	LogUnif(d_s,20)

We performed inference under a two-type birth-death model in which the two types correspond to the two host species. Following Kühnert *et al.* (2016) we used the following epidemiological parameterization of the multi-type birth-death model. Assuming lineage removal on sampling (i.e. $r_i = 1$), this parameterization can be expressed in the following way:

$$R_{ij} = \frac{\lambda_{ij}}{\mu_i + \psi_i}$$

$$\delta_i = \mu_i + \psi_i$$

$$s_i = \frac{\psi_i}{\mu_i + \psi_i}$$

Here diagonal elements of R_{ij} represent the host-specific effective reproductive number, while off-diagonal elements generalize this quantity to describe cross-type (here zoonotic) transmission events. The parameter δ_i is the so-called "become uninfectious" rate, which is the net rate at which infected hosts are removed from their respective infectious compartments.

All parameters were assumed constant through time, with the single exception of the host-specific sampling proportions s_i, which were assumed to be zero for all times outside of the interval between the first and last included sample from each host type. Prior distributions are shown in Table 1.

For the MCMC portion of the analysis, 5 independent chains of approximately 9×10^7 iterations each were run, which was sufficiently long that all parameters of interest had an effective sample size (ESS) greater than 200. These chains were then compared to assess statistical convergence. (Supplemental Figure S2 shows close agreement between sampled distributions from independent chains, indicating convergence.) The first 10% of iterations from each chain were discarded to account for burn-in, and these truncated chains were then combined for further analysis. Subsequently, stochastic mapping of ancestral types and particle filtering of epidemic trajectories were conducted on one in every 10^6 MCMC iterations,

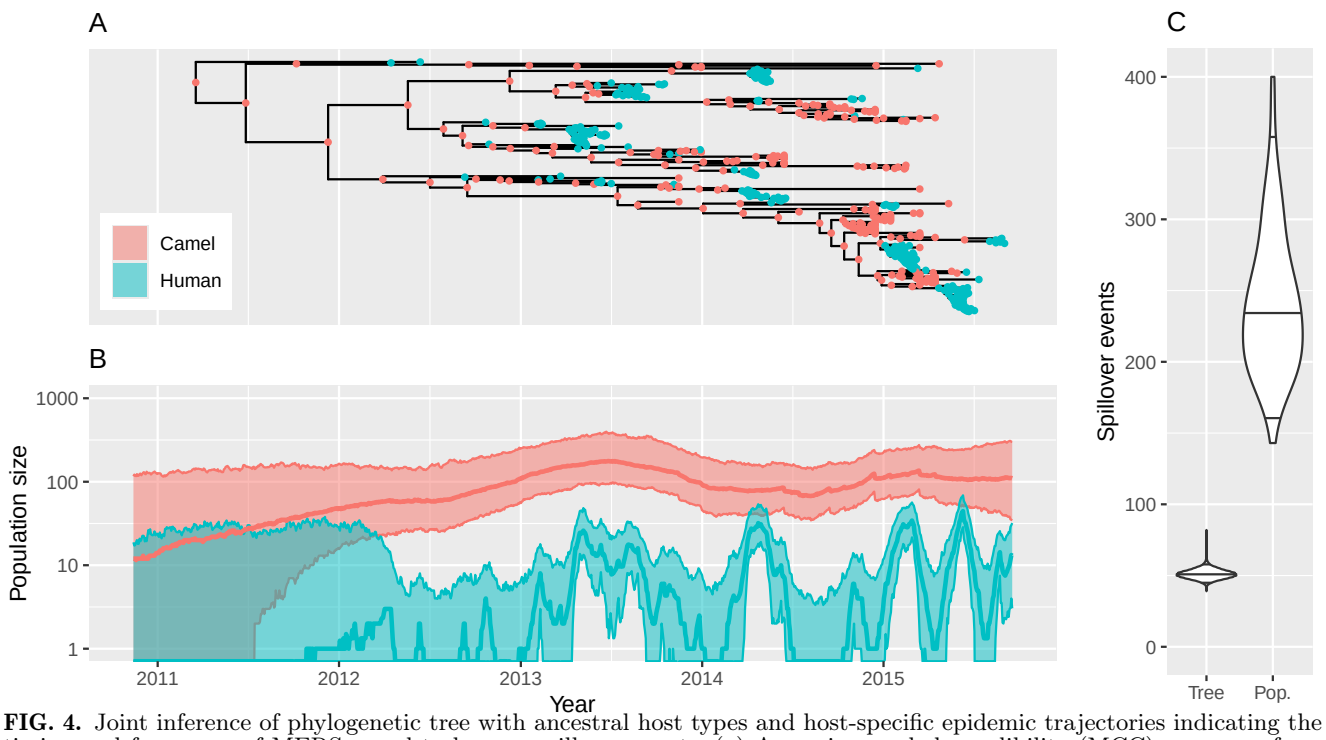

FIG. 4. Joint inference of phylogenetic tree with ancestral host types and host-specific epidemic trajectories indicating the timing and frequency of MERS camel to human spillover events. (a) A maximum clade credibility (MCC) summary from the dated phylogeny posterior with most highly probable host types at internal nodes marked. (b) Marginal posterior of infected host population size dynamics, with the thick central lines representing the median and the bounds of the shaded areas representing the 2.5% and 97.5% quantiles of the marginal host-specific infected population size posterior at each time point. (c) Marginal posteriors of spillover events present on the tree (i.e. ancestral to included samples) and present in the larger population since 2011.

yielding 410 sampled edge-typed trees and multi-type trajectories in total. Due to the large interval between MCMC iterations, we can regard these as effectively independent draws from the respective marginal tree and trajectory posteriors and thus sufficient to characterise these posteriors.

Figure 4 summarizes the results of this analysis. In Figure 4A, a maximum clade credibility (MCC) summary tree is shown with annotations indicating the ancestral host types best supported by the posterior edge-typed tree posterior. While minor differences exist, the same patterns observed by Dudas *et al.* (2018) using the structured coalescent model are also visible on this tree, with clear indication of a persistent camel-associated reservoir producing periodic spillovers into the human population.

By counting events on the edge-typed trees, we found that the distribution of camel to human transmissions directly ancestral to the dataset has a 95% HPD of [45,57] and median of 51 transmissions. These numbers are comparable to those determined by Dudas *et al.* (2018) applying the same approach to their structured coalescent results, who reported a 95% HPD of $[48-63]$ and a median of 56 for the number of transmissions.

The population-level spillover dynamics are shown in figure 4B, which displays a summary of the host-specific infected population size trajectory posteriors, including median values and 95% central posterior density intervals. The nearly constant case load in the camel population is clearly visible. (Interestingly,

this displays strong similarities to the Bayesian Skyline Plot estimate of the camel-associated effective population size shown figure 5 of Dudas *et al.* (2018).) Additionally, the spillover outbreaks in the human population are reflected in the well-resolved spikes in the human-associated population trajectory.

Importantly, the sampled trajectories can be regarded as sequences of model events. We used these to characterise the marginal probability distributions for the number of camel to human transmission events in the whole population from which our sample was drawn. Unsurprisingly, we estimate that the number of these events is significantly higher than the number of ancestral events, which the marginal posterior having a 95% highest posterior distribution (HPD) interval of [143,421] and a median of 236. A comparison of the marginal posterior distribution for the camel to human transmission events in the edge-typed tree with the corresponding distribution for events in the wider population is shown in figure 4C.

Discussion

Multi-type birth-death phylodynamics inference methods have traditionally focused on the inference of the number, type and timing of events, such as lineage divergences and type-change events, directly affecting lineages ancestral to sampled sequences or individuals. By enabling estimation of type-specific population size dynamics, the method presented in this paper extends this capability to allow direct inference of the number, type and timing of events in the larger population which are not necessarily directly ancestral to the sampled genetic data. The approach is applicable to structured (multi-type) birth-death models for which thus far inference of population size dynamics was not generally possible. (We note that the restriction of the model space to linear models—which cannot, for instance, directly account for density-dependent shifts in transmission rate—can be somewhat ameliorated through the application of time inhomogeneous rate variation, which our approach supports via arbitrary piecewise constant variation in rate functions, as shown by Kühnert *et al.* (2014).) This follows on from our previous work (Vaughan *et al.*, 2019) which used a particle marginal Metropolis-Hastings (PMMH) approach of Andrieu *et al.* (2010) to jointly sample both trees and population trajectory posteriors under a variety of linear and nonlinear birth-death models.

As demonstrated in the MERS-CoV case study, the ability to impute these latent multi-type population trajectory variables has clear applications to the study of large epidemics, where the specific details of the subset of the population ancestral to the sample set is often not of direct interest. Instead it is usually the larger encompassing population of infected hosts that we wish to probe. Beyond epidemiology, there are also natural applications in other fields which make use of birth-death phylodynamic models. In paleontology, models such as the FBD together with linear multi-type models—BiSSE (Maddison *et al.*,

2007), MuSSE (FitzJohn, 2012), HiSSE (Beaulieu and O'Meara, 2016), etc.—are commonly used to infer trait-dependent speciation and extinction rates from trees inferred using genetic and morphological data. Our approach provides an efficient and self-consistent route to estimating and testing hypotheses relating to the corresponding trait-dependent species richness dynamics under these same models.

Given the vast number of variables contained in the trajectories our approach aims to infer, which include the type and timing of every demographic event in the larger population, questions regarding the identifiability of these variables given limited data naturally arise. Our response to this is as follows. Firstly, while such questions are interesting and should be investigated further, we remind the reader that identifiability of all of the random variables in a particular analysis is less crucial in a Bayesian setting than it is when employing maximum likelihood methods. In place of a point estimate, Bayesian results are expressed in terms of a posterior probability distribution. Unidentifiability of a certain parameter leads only to posterior uncertainty for that parameter, with the regularizing influence of the prior ensuring the posterior remains proper. Secondly, it is clear from our results that genetic data provide useful signal for at least some elements of the multi-type trajectory event sequences, meaning that at least some useful parameters are identifiable. Finally, when the goal of a given phylodynamic analysis is to learn about properties of the larger population (i.e. beyond only the ancestors of the sampled individuals) we believe it is preferable to draw conclusions from posterior distributions which explicitly describe the uncertainty in these properties—even when this uncertainty is large. The alternative approach, namely extrapolating from narrower posterior distributions over proxy statistics specific to the subset of individuals ancestral to the sample set, is likely to result both bias and under-representation of the true uncertainty in the conclusions.

Even when the ancestral population or tree is the sole focus, inferred trajectories may still be useful in gauging the appropriateness of phylodynamic models. After all, birth-death-sampling processes imply the existence of such trajectories even when traditional methods report only the marginal distributions of the phylogenetic trees and parameters. Nonsensical properties of inferred trajectories may reveal existing but previously-obscured deficiencies in the ability of the model to adequately describe the intended population dynamic and sampling processes.

To conclude, we have presented a Bayesian approach for the inference of complete multi-type birth-death population size trajectories from genetic sequences. Our application of the method to MERS-CoV demonstrates that this can be used to directly infer interesting properties of populations from which genetic samples are taken, such as the host-specific infected population dynamics and the times and numbers of zoonotic events. Importantly, we have shown that extending existing MCMC-based

phylodynamic analyses to incorporate trajectory inference can be accomplished with negligible additional computational overhead. For these reasons, we believe that this new approach will prove to be a useful and practical addition to the phylodynamics tool set.

Acknowledgements

The authors gratefully acknowledge Nicola Mulberry who provided valuable feedback on an early version of the manuscript. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement no. 101001077.

Data availability

All of the BEAST 2 input files and associated scripts required to reproduce the analyses and results presented in this paper, together with corresponding usage instructions, are available for download from GitHub (`https://github.com/tgvaughan/MultiTypeTrajectoryAnalyses`). The method itself is implemented as a new package, BDMM-Prime, which is distributed as free software under the terms of the GNU General Public License version 3. The source code for the software is available from its GitHub repository (`https://github.com/tgvaughan/BDMM-Prime`), and a complete online manual including a detailed tutorial can be found at the package website (`https://tgvaughan.github.io/BDMM-Prime`).

References

Andrieu, C., Doucet, A., and Holenstein, R. 2010. Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3): 269–342.

Barido-Sottani, J., Vaughan, T. G., and Stadler, T. 2020. A multitype birth–death model for bayesian inference of lineage-specific birth and death rates. *Systematic Biology*, 69(5): 973–986.

Beaulieu, J. M. and O'Meara, B. C. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Systematic Biology*, 65(4): 583–601.

Beerli, P. and Felsenstein, J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A*, 98(8): 4563–4568.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N. D., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., and Drummond, A. J. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, 15(4): e1006650.

Cappe, O., Godsill, S. J., and Moulines, E. 2007. An overview of existing methods and recent advances in sequential monte carlo. *Proceedings of the IEEE*, 95(5): 899–924.

De Maio, N., Wu, C.-H., O'Reilly, K. M., and Wilson, D. 2015. New routes to phylogeography: A bayesian structured coalescent approximation. *PLoS Genet*, 11(8): e1005421.

de Wit, E., van Doremalen, N., Falzarano, D., and Munster, V. J. 2016. SARS and MERS: recent insights into emerging coronaviruses. *Nature Reviews Microbiology*, 14(8): 523–534.

Douglas, J., Zhang, R., and Bouckaert, R. 2021. Adaptive dating and fast proposals: Revisiting the phylogenetic relaxed clock model. *PLOS Computational Biology*, 17(2): e1008322.

Dudas, G., Carvalho, L. M., Rambaut, A., and Bedford, T. 2018. MERS-CoV spillover at the camel-human interface. *eLife*, 7.

Ewing, G., Nicholls, G., and Rodrigo, A. 2004. Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. *Genetics*, 168(4): 2407–2420.

FitzJohn, R. G. 2012. Diversitree : comparative phylogenetic analyses of diversification in r. *Methods in Ecology and Evolution*, 3(6): 1084–1092.

Freyman, W. A. and Höhna, S. 2018. Stochastic character mapping of state-dependent diversification reveals the tempo of evolutionary decline in self-compatible onagraceae lineages. *Systematic Biology*, 68(3): 505–519.

Gavryushkina, A., Heath, T. A., Ksepka, D. T., Stadler, T., Welch, D., and Drummond, A. J. 2017. Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Systematic biology*, 66(1): 57–73.

Gillespie, D. T. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.*, 22: 403.

Grear, D. A., Hall, J. S., Dusek, R. J., and Ip, H. S. 2017. Inferring epidemiologic dynamics from viral evolution: 2014–2015 eurasian/north american highly pathogenic avian influenza viruses exceed transmission threshold, r0 = 1, in wild birds and poultry in north america. *Evolutionary Applications*, 11(4): 547–557.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., and Holmes, E. C. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656): 327–32.

Guinat, C., Valenzuela Agüí, C., Vaughan, T. G., Scire, J., Pohlmann, A., Staubach, C., King, J., Świętoń, E., Dán, Á., Černíková, L., Ducatez, M. F., and Stadler, T. 2022. Disentangling the role of poultry farms and wild birds in the spread of highly pathogenic avian influenza virus in europe. *Virus Evolution*, 8(2).

Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2): 160–174.

Hey, J. and Nielsen, R. 2007. Integration within the felsenstein equation for improved markov chain monte carlo methods in population genetics. *Proc Natl Acad Sci U S A*, 104(8): 2785–2790.

Kendall, D. G. 1948. On the generalized "birth-and-death" process. *The Annals of Mathematical Statistics*, 19(1): 1–15.

King, A. A., Lin, Q., and Ionides, E. L. 2022. Markov genealogy processes. *Theoretical Population Biology*, 143: 77–91.

King, A. A., Lin, Q., and Ionides, E. L. 2024. Exact phylodynamic likelihood via structured markov genealogy processes. DOI:10.48550/arXiv.2405.17032. (preprint).

Kühnert, D., Stadler, T., Vaughan, T. G., and Drummond, A. J. 2014. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death sir model. *J R Soc Interface*, 11(94): 20131106.

Kühnert, D., Stadler, T., Vaughan, T. G., and Drummond, A. J. 2016. Phylodynamics with migration: A computational framework to quantify population structure from genomic data. *Molecular biology and evolution*, 33: 2102–2116.

Loiseau, C., Windels, E. M., Gygli, S. M., Jugheli, L., Maghradze, N., Brites, D., Ross, A., Goig, G., Reinhard, M., Borrell, S., Trauner, A., Dötsch, A., Aspindzelashvili, R., Denes, R., Reither, K., Beisel, C., Tukvadze, N., Avaliani, Z., Stadler, T., and Gagneux, S. 2023. The relative transmission fitness of multidrug-resistant mycobacterium tuberculosis in a drug resistance hotspot. *Nature Communications*, 14(1).

Maddison, W. P., Midford, P. E., and Otto, S. P. 2007. Estimating a binary character's effect on speciation and extinction. *Syst Biol*, 56(5): 701–710.

Manceau, M., Gupta, A., Vaughan, T., and Stadler, T. 2021. The probability distribution of the ancestral population size conditioned on the reconstructed phylogenetic tree with occurrence data. *Journal of Theoretical Biology*, 509: 110400.

Müller, N. F., Rasmussen, D., and Stadler, T. 2018. MASCOT: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics*, 34(22): 3843–3848.

Nadeau, S. A., Vaughan, T. G., Scire, J., Huisman, J. S., and Stadler, T. 2021. The origin and early spread of sars-cov-2 in europe. *Proceedings of the National Academy of Sciences*, 118(9).

Rabosky, D. L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS ONE*, 9(2): e89543.

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

Scire, J., Barido-Sottani, J., Kühnert, D., Vaughan, T. G., and Stadler, T. 2022. Robust phylodynamic analysis of genetic sequencing data from structured populations. *Viruses*, 14(8): 1648.

Stadler, T. 2010. Sampling-through-time in birth-death trees. *J Theor Biol*, 267(3): 396–404.

Stadler, T., Kühnert, D., Bonhoeffer, S., and Drummond, A. J. 2012. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci U S A*, 110(1): 228–233.

Stadler, T., Pybus, O. G., and Stumpf, M. P. H. 2021. Phylodynamics for cell biologists. *Science*, 371(6526).

Vaughan, T. G., Kühnert, D., Popinga, A., Welch, D., and Drummond, A. J. 2014. Efficient bayesian inference under the structured coalescent. *Bioinformatics*, 30(16): 2272–2279.

Vaughan, T. G., Leventhal, G. E., Rasmussen, D. A., Drummond, A. J., Welch, D., and Stadler, T. 2019. Estimating epidemic incidence and prevalence from genomic data. *Molecular Biology and Evolution*, 36: 1804–1816.

World Health Organization 2022. Middle East respiratory syndrome coronavirus (MERS-CoV) (Electronic resource, https://www.who.int/news-room/fact-sheets/detail/middle-east-respiratory-syndrome-coronavirus-(mers-cov), last accessed 27 August 2024), Geneva, Switzerland.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *Journal of molecular evolution*, 39: 306–314.

Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D., and Fouchier, R. A. 2012. Isolation of a novel coronavirus from a man with pneumonia in saudi arabia. *New England Journal of Medicine*, 367(19): 1814–1820.

Zhang, Z., Shen, L., and Gu, X. 2016. Evolutionary dynamics of mers-cov: Potential recombination, positive selection and transmission. *Scientific Reports*, 6(1).