Title

Data Integration in Bayesian Phylogenetics

Journal

Annual Review of Statistics and Its Application, 10(1)

ISSN

2326-8298

Authors

Hassler, Gabriel W
Magee, Andrew
Zhang, Zhenyu
et al.

Publication Date

2023-03-10

DOI

10.1146/annurev-statistics-033021-112532

Peer reviewed

HHS Public Access

Author manuscript

Annu Rev Stat Appl. Author manuscript; available in PMC 2024 May 21.

Published in final edited form as:

Annu Rev Stat Appl. 2023 ; 10: 353–377. doi:10.1146/annurev-statistics-033021-112532.

Data integration in Bayesian phylogenetics

Gabriel W Hassler[1], **Andrew Magee**[2], **Zhenyu Zhang**[2], **Guy Baele**[3], **Philippe Lemey**[3], **Xiang Ji**[4], **Mathieu Fourment**[5], **Marc A Suchard**[1,2,6]

[1]Department of Computational Medicine, University of California, Los Angeles, USA, 90095

[2]Department of Biostatistics, University of California, Los Angeles, USA, 90095

[3]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium, 3000

[4]Department of Mathematics, Tulane University, New Orleans, USA, 70118

[5]Australian Institute for Microbiology and Infection, University of Technology Sydney, Ultimo NSW, Australia, 2007

[6]Department of Human Genetics, University of California, Los Angeles, USA, 90095

Abstract

Researchers studying the evolution of viral pathogens and other organisms increasingly encounter and use large and complex data sets from multiple different sources. Statistical research in Bayesian phylogenetics has risen to this challenge. Researchers use phylogenetics not only to reconstruct the evolutionary history of a group of organisms, but also to understand the processes that guide its evolution and spread through space and time. To this end, it is now the norm to integrate numerous sources of data. For example, epidemiologists studying the spread of a virus through a region incorporate data including genetic sequences (e.g. DNA), time, location (both continuous and discrete) and environmental covariates (e.g. social connectivity between regions) into a coherent statistical model. Evolutionary biologists routinely do the same with genetic sequences, location, time, fossil and modern phenotypes, and ecological covariates. These complex, hierarchical models readily accommodate both discrete and continuous data and have enormous combined discrete/continuous parameter spaces including, at a minimum, phylogenetic tree topologies and branch lengths. The increased size and complexity of these statistical models have spurred advances in computational methods to make them tractable. We discuss both the modeling and computational advances below, as well as unsolved problems and areas of active research.

Keywords

Bayesian networks; continuous-time Markov processes; Gaussian processes; phylogenetic comparative methods; phylogeography

ghassler@ucla.edu .

1. Introduction

All living things on the planet share a common evolutionary history. Phylogenetic trees capture the evolutionary relationships between groups of organisms (Baldauf 2003). At the extremes, these phylogenies can describe the evolution of all life on earth spanning ~ 4 billion years or that of a viral lineage over weeks. Statistical phylogenetics gives researchers the tools to study these evolutionary processes and can be used to answer both fundamental biological questions, such as "which species of ape is most closely related to humans and when did our evolutionary histories diverge?" (Bradley 2008) and more practical ones such as "how effective are various interventions at controlling the spread of a viral epidemic?" (Dellicour et al. 2018). Researchers typically rely on molecular sequences (e.g. DNA, RNA, amino acids) to infer the phylogeny itself and commonly incorporate additional sources of data to answer specific questions. For example, toward the end of this review in Section 4 we examine a case study where researchers investigate the early spread of SARS-CoV-2, the virus that causes COVID-19, across the world (Lemey et al. 2020). This analysis incorporates viral genetic sequences, sample collection dates and locations, individual-level travel history, global air traffic patterns, local SARS-CoV-2 case counts and within-host infection dynamics into a coherent statistical model that allows researchers to reconstruct the early pathways along which SARS-CoV-2 spread early in the pandemic.

From a statistical perspective, phylogenetics offers a rich array of complex hierarchical models for both inferring the phylogeny itself as well as parameters associated with the underlying evolutionary processes of interest (Nascimento et al. 2017). The complexity of these models, however, can result in theoretical and computational challenges to inference that limit their scalability. These challenges have led to the development of statistical methods with broad utility beyond the field of phylogenetics itself. In this review, we first introduce the fundamental statistical approaches to phylogenetics in Section 1.1 and the advantages of the Bayesian approach in Section 1.2 below. We then discuss modern methods for inferring phylogenetic trees in Section 2 and data integration in Section 3. As mentioned previously, we examine in Section 4 a case study that relies on many of the methods discussed in earlier sections.

1.1. Molecular evolution on a phylogenetic tree

Let the phylogenetic tree $\mathcal{T}$ be a bifurcating directed acyclic graph with N degree-one terminal/tip nodes $v_1, \ldots, v_N$, $N-2$ degree-three internal nodes $v_{N+1}, \ldots, v_{2N-2}$ and one degree-two root node v_{2N-1}. With the exception of the root node, there is an edge connecting each node v_i to its parent $v_{\text{pa}(i)}$ with length t_i. See Figure 1 for a simple example. Depending on the statistical model, these edge lengths are typically proportional to either the amount of time or expected number of genetic changes separating nodes v_i and $v_{\text{pa}(i)}$. While some parameterizations permit multifurcations/polytomies (i.e. nodes with more than two children), we focus on the bifurcating case without loss of generality as multifurcations can be represented via bifurcations with edge lengths equal to zero. Note that some parameterizations assume unrooted trees where the degree-two root node is omitted. In the unrooted case, the phylogeny is no longer directed and there are no fixed parent/child relationships between nodes.

Likelihood-based phylogenetic inference typically relies on molecular sequences **S** to inform the phylogenetic tree. The tree $\mathscr{F}$ parameter space is divided into a discrete topology space (i.e. the bifurcating tree structure without the edge lengths) and a continuous edge length space. The edge lengths inhabit a (non-negative) continuous $(2N-2)$−dimensional space, $(t_1, \ldots, t_{2N-2}) \in \mathbb{R}^{2N-2}_{\geq 0} = \{(x_1, \ldots, x_{2N-2}) : x_i \geq 0\}$. The space of tree topologies is unordered, discrete, and grows combinatorially in the number of tips, with $(2N-3)!! = \prod_{i=1}^{N-1} 2i-1$ possible tree topologies for N tips.

There are many ways to specify the likelihood $p(\mathbf{S} \mid \mathscr{F})$ that are beyond the scope of this review (see Felsenstein (2004), Sullivan & Joyce (2005), Lemey et al. (2009b) for overviews). However, it is useful to sketch a common form of these likelihoods. Let us assume that we have DNA characters, comprising the nucleotides A, C, G and T (the building blocks of DNA). We make the standard assumption that the molecular sequences **S** are aligned into an $N \times M$ matrix, where M is the number of nucleotides in a sequence alignment. Each column, called a site, in this alignment represents a homology assumption, in that all characters in a column share a single common ancestor somewhere back in time. We also commonly assume that each site evolves independently and identically (with the other sites) along the tree according to a four-state continuous-time Markov process with the instantaneous rate matrix **Q**. Let s_i^m be the nucleotide at site m for node v_i. The transition probability of observing s_i^m given the parent nucleotide state $s_{\mathrm{pa}(i)}^m$ and edge length t_i is $p_{s_i^m s_{\mathrm{pa}(i)}^m}$, such that $\mathbf{P} = \{p_{\ell m}\} = \exp(t_i \mathbf{Q})$ forms the transition probability matrix.

The clear challenge to computing likelihoods under this model is that we have not observed any sequence data associated with the internal nodes $v_{N+1}, \ldots, v_{2N-2}$ or the root node v_{2N-1} and so must marginalize over their values. Assuming independence between sites and a prior $p(s_{2N-1}^m)$ on the root, the likelihood can then be expressed as

$$p(\mathbf{S} \mid \mathscr{F}) = \prod_{m=1}^{M} \sum_{s_{N+1}^m \in \{A,C,G,T\}} \cdots \sum_{s_{2N-1}^m \in \{A,C,G,T\}} p(s_{2N-1}^m) \prod_{i=1}^{2N-2} p(s_i^m \mid s_{\mathrm{pa}(i)}^m, t_i). \quad 1.$$

Naive computation of the above equation requires summing over 4^{N-1} unobserved states and is computationally intractable. Felsenstein's pruning algorithm (Felsenstein 1973a, 1981), however, uses a post-order traversal of the tree to compute this likelihood in $\mathcal{O}(N)$ time, and all modern implementations of this likelihood calculation rely on that basic approach. The fundamental approach of this pruning algorithm is based on dynamic programming and has found repeated rediscovery in the message-passing algorithm (Pearl 1982) and sum-product algorithm (Kschischang et al. 2001).

Let $\mathbf{s}^m$ be the nucleotides at site m associated with all tip nodes. The pruning algorithm relies on recursively computing the probability mass function $p(\mathbf{s}_{[i]}^m \mid s_i^m, \mathscr{F}_{[i]})$, where $\mathscr{F}_{[i]}$ is the sub-tree with root node v_i, and $\mathbf{s}_{[i]}^m$ is the sub-vector of $\mathbf{s}^m$ restricted to the tips in

$\mathscr{T}_{\lfloor i \rfloor}$. At the root node v_{2N-1}, $\mathscr{T}_{\lfloor i \rfloor} = \mathscr{T}$ and $\mathbf{s}^m_{\lfloor 2N-1 \rfloor} = \mathbf{s}^m$, and the pruning algorithm computes $p(\mathbf{s}^m \mid s^m_{2N-1}, \mathscr{T}) = p(\mathbf{s}^m_{\lfloor 2N-1 \rfloor} \mid s^m_{2N-1}, \mathscr{T}_{\lfloor 2N-1 \rfloor})$ via the following recursive relationship:

$$\begin{aligned} p(\mathbf{s}^m_{\lfloor i \rfloor} \mid s^m_i, \mathscr{T}_{\lfloor i \rfloor}) &= p(\mathbf{s}^m_{\lfloor j \rfloor} \mid s^m_i, \mathscr{T}_{\lfloor i \rfloor}) p(\mathbf{s}^m_{\lfloor k \rfloor} \mid s^m_i, \mathscr{T}_{\lfloor i \rfloor}) \\ &= \sum_{s^m_j \in \{A,C,G,T\}} p(\mathbf{s}^m_{\lfloor j \rfloor} \mid s^m_j, \mathscr{T}_{\lfloor j \rfloor}) p(s^m_j \mid s^m_i, t_j) \\ &\times \sum_{s^m_k \in \{A,C,G,T\}} p(\mathbf{s}^m_{\lfloor k \rfloor} \mid s^m_k, \mathscr{T}_{\lfloor k \rfloor}) p(s^m_k \mid s^m_i, t_k), \end{aligned} \quad 2.$$

where nodes v_j and v_k are the children of node v_i. When the recursion reaches tip nodes $i = 1, \ldots, N$, $p(\mathbf{s}^m_{\lfloor i \rfloor} \mid s^m_i, \mathscr{T}_{\lfloor i \rfloor}) = 1_{\{\mathbf{s}^m_{\lfloor i \rfloor} = s^m_i\}}$, and the actual computations of computing the likelihood are performed via a post-order traversal of the tree (i.e. tips to root). The algorithm marginalizing over the root sequences

$$p(\mathbf{s}^m \mid \mathscr{T}) = \sum_{s^m_{2N-1} \in \{A,C,G,T\}} p(\mathbf{s}^m \mid s^m_{2N-1}, \mathscr{T}) p(s^m_{2N-1}) \quad 3.$$

and calculating $p(\mathbf{S} \mid \mathscr{T}) = \prod_{m=1}^{M} p(\mathbf{s}^m \mid \mathscr{T})$ is shown in Figure 2 on a simple example.

1.2. Why Bayesian?

In Bayesian phylogenetic inference, a common goal is to compute the posterior distribution of the phylogenetic tree given our sequence data,

$$p(\mathscr{T} \mid \mathbf{S}) \propto p(\mathbf{S} \mid \mathscr{T}) p(\mathscr{T}). \quad 4.$$

The tree prior $p(\mathscr{T})$ typically falls into one of two biologically-motivated families. Coalescent models (Kingman 1982, Strimmer & Pybus 2001, Minin et al. 2008, Müller et al. 2017, Faulkner et al. 2020) are based on population genetic abstractions of sampling a (relatively) small number of sequences from a large population. Birth-death models (Thompson et al. 1975, Nee et al. 1994, Stadler 2010, Höhna et al. 2019, Barido-Sottani et al. 2020, MacPherson et al. 2022) provide a forward-in-time model for the origination and termination of entire lineages. Bayesian approaches offer several advantages which we discuss below.

1.2.1. Quantifying uncertainty.—

Bayesian phylogenetics grew largely from the need to quantify and accommodate uncertainty in the phylogenetic tree (Sinsheimer et al. 1996, Rannala & Yang 1996). Measuring uncertainty in the phylogenetic tree is a fundamentally challenging problem as the primary parameter of interest is often the tree topology: a high-dimensional, unordered, tip-labeled discrete parameter. Typical uncertainty estimates focus on estimating the statistical support for a specific monophyletic clade (i.e. a group of taxa comprising all the descendants of a given ancestor). Prior to the advent of Bayesian

phylogenetic inference, phylogenetic uncertainty had been addressed with non-parametric bootstrapping (Felsenstein 1985a) with much confusion as to interpretation of the bootstrap *p*-value (see Hillis & Bull 1993, Felsenstein & Kishino 1993, Efron et al. 1996, Berry & Gascuel 1996). Bayesian posterior probabilities provided both an intuitive and statistically coherent method of addressing this uncertainty (Alfaro et al. 2003).

1.2.2. Time-resolved trees.—Early phylogenetic models focused on the case where branch lengths are measured in genetic distances and thus unconstrained by time. However, Bayesian approaches can naturally accommodate the time-constrained case in a hierarchical model. As the bulk of the review assumes such models, we briefly consider the structure of a time-calibrated phylogenetic model. First, a tree arises from the tree prior $p(\mathcal{F})$. The branch lengths $t_1, \ldots, t_{2N-2}$ of $\mathcal{F}$ are in calendar time. For each branch is a branch rate θ_i, such that the probability of changes along the branch is given by $\exp(t_i\theta_i\mathbf{Q})$. The prior on all branch rates $p(\theta_1, \ldots, \theta_{2N-2})$ is known as the (molecular) clock model (Zuckerkandl & Pauling 1962). Clock models typically either assume all branch rates are independent and identically distributed (Drummond et al. 2006) or that rates themselves evolve along the tree according to a correlated process (Thorne et al. 1998, Drummond & Suchard 2010).

1.2.3. Tree as nuisance parameter.—Phylogenetic methods offer opportunities to do more than just reconstruct the evolutionary history of a group of organisms. The branching patterns in trees themselves can be informative about patterns and processes governing biodiversity, such as mass extinctions (Stadler 2011, May et al. 2016), or the rate of spread of infectious diseases (Stadler et al. 2012, 2013). When combined with other information, such as the locality of samples or evolutionary traits, phylogenetic models provide a powerful framework for studying the spatiotemporal spread of both species and diseases, as well as the evolution of important traits (see Section 3). In many such cases, the tree itself is a nuisance parameter. Bayesian inference via Markov chain Monte Carlo (MCMC) provides a natural approach to numerically marginalize over the phylogenetic tree and study processes that condition on the tree independent of any single fixed tree's influence (Huelsenbeck et al. 2000, 2001, Suchard et al. 2001).

2. Modern phylogenetics: big trees and complex models

Early practitioners of Bayesian phylogenetics naturally used MCMC to sample from the posterior distribution of phylogenetic trees. Since it is relatively straightforward to marginalize over continuous nuisance parameters (e.g. the molecular substitution rate matrix **Q**), attention quickly turned to improving the efficiency with which the Markov chain explores tree space (Yang & Rannala 1997, Larget & Simon 1999, Mau et al. 1999, Li et al. 2000, Huelsenbeck & Ronquist 2001). This in turn gave rise to the observation that navigating tree space is hard (Lakner et al. 2008, Höhna & Drummond 2012, Whidden & Matsen IV 2015, Harrington et al. 2021).

We explore several solutions to this problem below. In Section 2.1, we discuss approaches to improving the efficiency of MCMC-based methods. We then discuss in Section 2.2 alternatives to MCMC inspired by phylogenetic problems. As these approaches permit

researchers to more efficiently explore the space of phylogenetic trees, we revisit in Section 2.3 the problem of assessing uncertainty in the phylogeny estimates.

2.1. MCMC-based approaches

MCMC is the workhorse of Bayesian phylogenetic inference. The efficiency of MCMC depends on two factors: the auto-correlation between parameter proposals and the speed at which proposals are made and evaluated. Researchers have relied on and contributed to numerous innovative computational and statistical methods in search of MCMC approaches that efficiently explore the high-dimensional tree space.

2.1.1. Faster likelihood calculations.—In the absence of known conjugate priors, efficient likelihood calculations are critical for efficient MCMC. As common models of sequence evolution assume conditional independence between different sites in the genome, parallelization is a natural approach toward fast computation. The BEAGLE (Suchard & Rambaut 2009, Ayres et al. 2012, 2019) and PLL (Izquierdo-Carrasco et al. 2013, Flouri et al. 2015) libraries leverage the computational power of multi-core processors, including graphics processing units (GPUs) in the former case, to massively parallelize likelihood calculations and accelerate computation. These libraries also cache calculations on sub-trees such that unnecessary calculations are not repeated when, for example, a branch length on one part of the tree is updated that does not influence the partial likelihood of other parts of the tree.

2.1.2. Sampling from high-dimensional posterior distributions.—The dimensionality of many continuous parameters (e.g. the branch lengths) scales with the size of the phylogenetic tree. Phylogenetic analyses commonly partition genetic sequences into different genes (or some other genetic unit) that evolve independently conditional on a tree. Modern Bayesian phylogenetic analyses include trees with thousands of tips (e.g. Lemey et al. 2021) and, as such, require inference of the joint posterior of thousands of highly-correlated parameters.

Baele et al. (2017a) develop an adaptive Metropolis (AM) algorithm (Haario et al. 2001) that leverages the parallel computing to take advantage of the conditional independence of the genetic partitions. The AM algorithm is a modification of MCMC where proposal distributions are informed by the empirical posterior distribution up to that point in the chain. While AM is non-Markovian, it remains ergodic under weak assumptions (Roberts & Rosenthal 2009). Baele et al. (2017a) update the chain via partition-specific multivariate Gaussian proposals with covariance influenced by the empirical posterior covariance of relevant parameters. The conditionally independent parameter blocks allow parallel likelihood computations, and the multivariate Gaussian proposals informed by the posterior have higher acceptance probability than naive multivariate proposals.

Hamiltonian Monte Carlo (HMC) is now a standard tool across Bayesian statistics for sampling from high-dimensional posterior distributions. At its core, HMC also uses information about the posterior to generate high-dimensional parameter proposals with high acceptance probability. As the aforementioned information originates from the gradient of the log-posterior with respect to the parameters of interest, efficient gradient calculations

are essential for efficient HMC. Ji et al. (2020) develop an $\mathcal{O}(N)$ algorithm for computing the gradient of the log-posterior with respect to all branch lengths simultaneously. These gradient calculations are also parallelizable using existing libraries (see Section 2.1.1) and result in an order of magnitude increase in computational efficiency.

2.1.3. Navigating tree space.—The discrete tree topology with $(2N - 3)!!$ possible states is often the most difficult model parameter to efficiently sample. As many other parameters, including the branch lengths and latent data associated with internal nodes, are only identifiable in the context of a particular tree, MCMC proposals that make large changes to the tree topology frequently have very low acceptance probability.

HMC is a standard tool for sampling from high-dimensional, highly correlated, continuous parameter spaces, but the discrete, combinatorial nature of the tree topology does not permit traditional HMC approaches. Dinh et al. (2017) develop probabilistic path HMC (PPHMC) to sample from spaces that form an orthant complex. Essentially, they sample the branch lengths via HMC in a way that branch lengths may approach 0. When HMC causes a branch length to cross 0, PPHMC randomly selects from one of the three equivalent topologies resulting from the zero branch length. To reduce error from the leapfrog approximation crossing non-differentiable orthant boundaries, they introduce a smoothing function at these boundaries, which dramatically increases the accuracy of the approximation of the Hamiltonian trajectory and Metropolis-Hastings acceptance probability. Similar work outside of the phylogenetic context includes that of Pakman & Paninski (2013), Mohasel Afshar & Domke (2015) and Nishimura et al. (2020).

More recently, Meyer (2021) has developed a series of AM procedures for efficiently navigating the space of unrooted tree topologies. Like other AM algorithms, these approaches rely on statistics of the posterior sample up to a point in a chain to inform future parameter proposals. In the context of tree topologies, the relevant statistics rely on the fact that each branch splits the taxa into two groups. The Meyer (2021) approach relies on the posterior frequency of these splits for each possible group of taxa, with topology proposals more likely to disrupt low-frequency splits than high-frequency splits. Similarly, Zhang et al. (2020) use parsimony (i.e. the minimum number of genetic changes necessary to account for the observed genetic diversity) to inform tree proposals, with highly parsimonious (i.e. few changes) proposals more likely than less parsimonious ones.

2.2. Beyond MCMC

2.2.1. Sequential Monte Carlo.—Teh et al. (2007) propose sequential Monte Carlo (SMC) for inferring tree-structured models. Due to the hierarchical structure of the model, the intermediate distributions are defined over forests (i.e. groups of sub-trees) over the observed sequences, and hence the dimension of the target distributions increases over each iteration. Based on this idea, Bouchard-Côté et al. (2012) propose an efficient framework, based on partially ordered set structures, which imposes restrictions on proposal distributions so that the final iteration results in valid phylogenetic trees. Since this phylogenetic SMC is restricted to jointly estimate tree topology and branch length distributions, Wang et al. (2015) propose particle MCMC which combines a combinatorial

SMC within an MCMC in order to jointly approximate other continuous parameters such as the parameters of the substitution rate matrix **Q**. Borrowing ideas from annealed importance sampling, Wang et al. (2020) put forward an annealed SMC algorithm to approximate the full phylogenetic model and, as other SMC-based methods, enable the computation of the marginal likelihood.

SMC has also been investigated in an online setting in which a posterior sample of trees is already available from a previous analysis (e.g. MCMC or SMC) and one wishes to directly update the posterior approximation with additional sequences. Dinh et al. (2017) show consistency of online SMCs in terms of weak convergence while Fourment et al. (2018) develop sophisticated proposals that better match the proposal density to the posterior.

2.2.2. Variational inference.—Until recently, variational inference (VI) has received limited attention in the field of phylogenetics, perhaps due to 1) the absence of conjugate prior distributions in the nearly all phylogenetic models and 2) the difficulty of analytically calculating the gradient of complex joint distributions. Dang & Kishino (2019) develop a computationally efficient VI-based method to approximate a model which allows different equilibrium frequencies across sequence sites. Since the likelihood of this model is in the exponential family, most of the expectations required for optimization are obtained in closed form. This method is restricted to unrooted trees and the authors used closed-form coordinate ascent and stochastic VI algorithms for solving the optimization problem. Fourment et al. (2020) use VI to approximate the marginal likelihood of fixed unrooted topologies using stochastic gradient ascent with analytical derivatives. Using the Stan language (Carpenter et al. 2017) and its automatic differentiation library, Fourment & Darling (2019) propose a framework for approximating complex models, including time-calibrated phylogenies with tree priors (e.g. coalescent models), molecular clock, and discrete phylogeography models.

The methods described so far only approximate continuous parameters of a fixed topology and therefore evade the combinatorial problem of the discrete topology space. The first approach developed to tackle this problem was introduced by Zhang & Matsen IV (2018a) using a general Bayesian network formulation for tree probability estimation. Given a set of topologies, this structure provides an accurate and rich distribution over the topology space. Subsequently, the same authors (Zhang & Matsen IV 2018b) build on the Bayesian network idea and propose jointly approximating the tree structure and the branch length distributions. This method also necessitates a set of topologies to define the structure of the Bayesian network, however dynamic construction of the network is an active area of research. Moretti et al. (2021) propose a hybrid method using VI and combinatorial SMC to approximate posteriors defined on the space of phylogenetic trees. The main advantage of this method is that it does not require precomputing a set of topologies. With the exception of the Stan-based method which allows approximating a posterior using a multivariate normal distribution, every method described so far uses meanfield approximation thereby ignoring correlation between parameters. Since parameters in phylogenetic models tend to be highly correlated, Zhang (2020) proposes to use normalizing flows to improve the expressiveness of the approximate distribution.

Recently, Ki & Terhorst (2022) synthesized this VI-based work with phylodynamic methods to fit a complex epidemiological model with thousands of sequences. The authors showed that their method was order of magnitude faster than an MCMC-based approaches and was able recover acceptable parameter estimates.

2.3. Uncertainty in tree space revisited

As discussed in Section 1.2.1, Bayesian phylogenetic methods conveniently quantify uncertainty in the tree. Many evolutionary questions can be phrased as "is there a subtree in the phylogeny which contains all of (some set of) sequences and no other sequences?" With MCMC samples in hand, we can easily obtain this probability by counting MCMC samples with the subtree. The fact that this estimate can carry substantial Monte Carlo error is often ignored. For continuous random variables, Monte Carlo error is typically addressed using the effective sample size (ESS, i.e. the number of independent samples which would yield the same standard error of the mean). Trees, however, are more complex objects.

Gaya et al. (2011) introduce one approach that focuses on taxa splits (i.e. bi-partitions of the tips by cutting the tree at a given edge). The tree is reduced to a series of indicator variables denoting whether a given split is present or absent in each tree. Uncertainty in the probability of specific splits can then be expressed via the ESS of these indicators. Fabreti & Höhna (2021) observe, however, that this approach has difficulty with splits whose probabilities approach 0 or 1. They also note that the Gaya et al. (2011) ESS incorrectly assumes that splits are independent. Regardless, Fabreti & Höhna (2021) find evidence via simulation that the Gaya et al. (2011) approach may remain robust.

Lanfear et al. (2016) propose an ESS for the phylogeny itself. They suggest two approaches based on distances between trees. One such approach is the pseudo-ESS, where for each posterior tree sample the distance is computed to all other tree samples. The overall tree ESS is taken to be the median of the ESSs of these distance metrics. Lanfear et al. (2016), however, do not establish any link between this pseudo-ESS and Monte Carlo error.

Magee et al. (2021) develop several additional approaches for computing the ESS of a phylogeny. One such approach employs Fréchet generalizations of covariance such that the generalized auto-correlation ρ_t between trees can be computed and the following standard identity can be applied: $\text{ESS} = n/(\sum_{t=-\infty}^{\infty} \rho_t)$. Additionally, Magee et al. (2021) propose a simulation-based approach to test whether a putative tree ESS is useful for quantifying Monte Carlo error in the tree. They find that most tested tree ESS measures can capture Monte Carlo error in the probabilities of splits, as well as other important summaries of the posterior distribution. The tree ESS approaches additionally do not appear to suffer from the difficulties Fabreti & Höhna (2021) identified with low and high probability splits.

3. Data Integration

In many cases the phylogenetic tree is actually a nuisance parameter and not of scientific interest itself (see Section 1.2.3). Rather, there is some other process (e.g. rate of viral transmission between two locations, strength of natural selection) that is separate from yet dependent on the evolutionary history that researchers would like to explore. In these

cases, researchers frequently seek to integrate varying sources of data into a single, coherent statistical model of evolution. These additional sources of data frequently include time (see Section 1.2.2) and geographic location (Lemey et al. 2009a, 2010).

Before discussing specific statistical models for integrating varying types of data, we first introduce a general framework in which to orient these models in Section 3.1. We then examine models and inference methods associated with integrating both discrete and continuous data into phylogenetic models in Sections 3.2 and 3.3, respectively. While we briefly discuss applications in the sections below, Baele et al. (2017b) offer a more thorough overview of the different kinds of data integrated into these phylogenetic models.

3.1. A unified modeling framework

There are myriad statistical models for integrating additional data into these phylogenetic models. While each model is naturally tailored to a specific application, most share a common, general framework (see Section 3.4.3 for a notable exception). Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{iK})^t$ be a vector of latent traits associated with node ν_i for $i = 1, \ldots, 2N - 1$. Similarly, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iP})$ be the data associated with tip nodes $\nu_1, \ldots, \nu_N$. For tips $i = 1, \ldots, N$, we posit a possibly stochastic link function $\mathbf{y}_i = f(\mathbf{x}_i)$.

These models describe a data generative process where the distribution of each $\mathbf{x}_i$ conditional on the trait values of its parent $\mathbf{x}_{\text{pa}(i)}$ are distributed with density or mass function $p\left(\mathbf{x}_i \mid \mathbf{x}_{\text{pa}(i)}\right) = g(\mathbf{x}_i; \mathbf{x}_{\text{pa}(i)}, \boldsymbol{\theta}_i, \boldsymbol{\Theta})$, where $\boldsymbol{\theta}_i$ represents branch-specific parameters, and $\boldsymbol{\Theta}$ represent universal model parameters. Typically, $\boldsymbol{\theta}_i$ includes at the very minimum the branch length t_i. By placing a prior on the root $p(\mathbf{x}_{2N-1} \mid \boldsymbol{\theta}_{2N-1})$, we can define a likelihood over the data $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)^t$:

$$p(\mathbf{Y} \mid f, g, \mathcal{T}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{2N-1}, \boldsymbol{\Theta}). \quad 5.$$

See Figure 3 for a model schematic.

While this framework seems (and indeed is) incredibly generalizable, all models resulting from it share a critical property: once lineages diverge they evolve independently. To formalize this notion, assume two nodes ν_i and ν_j that share a common parent $\nu_{\text{pa}(i)} = \nu_{\text{pa}(j)} = \nu_k$. Let $\mathbf{Y}_{|i|}$ and $\mathbf{Y}_{|j|}$ be the data associated with all tip nodes descended from node ν_i and ν_j, respectively. By construction, $\mathbf{Y}_{|i|} \mid \mathbf{x}_k$ and $\mathbf{Y}_{|j|} \mid \mathbf{x}_k$ are independent. This conditional independence is a defining feature of these phylogenetic models that statisticians routinely exploit to increase computational efficiency of statistical inference.

Readers may note that the model of molecular sequence evolution described in Section 1.1 fits neatly within this more general framework. Specifically, the data $\mathbf{Y}$ are comprised of discrete nucleotides (e.g. $y_{ij} \in \{A, C, G, T\}$), the link function $f(\mathbf{x}_i) = \mathbf{x}_i$, and the probability mass function $g(\mathbf{x}_i; \mathbf{x}_{\text{pa}(i)}, t_i, \mathbf{Q}) = \prod_{m=1}^{M} \exp(t_i \mathbf{Q})_{x_{\text{pa}(i)m} x_{im}}$.

As noted above, Bayesian methods (specifically MCMC) offer a to-date unmatched ability to study evolutionary processes without conditioning on an particular evolutionary history. This follows simply from the fact that researchers can easily sample from the marginal density of a parameter of interest from a realized MCMC simulation. Let $\mathbf{\Phi}$ represent all parameters associated with nucleotide evolution (e.g. the substitution rate matrix $\mathbf{Q}$) and let $\mathbf{\Psi} = \{\theta_1, \ldots, \theta_{2N-1}, \mathbf{\Theta}\}$ be the parameters associated with some separate trait-evolutionary process. One can then sample from the posterior

$$p(\mathscr{T}, \mathbf{\Phi}, \mathbf{\Psi} \mid \mathbf{S}, \mathbf{Y}) \propto p(\mathbf{S} \mid \mathscr{T}, \mathbf{\Phi}) p(\mathbf{Y} \mid \mathscr{T}, \mathbf{\Psi}) p(\mathscr{T}) p(\mathbf{\Phi}) p(\mathbf{\Psi}) \quad 6.$$

via a Metropolis-within-Gibbs approach (Gelfand 2000) where one iteratively samples from $p(\mathbf{\Phi} \mid \mathscr{T}, \mathbf{S})$, $p(\mathbf{\Psi} \mid \mathscr{T}, \mathbf{Y})$, and $p(\mathscr{T} \mid \mathbf{S}, \mathbf{Y}, \mathbf{\Phi}, \mathbf{\Psi})$. This compartmentalization of the inference procedure means that methods for sampling from the nucleotide substitution parameters $\mathbf{\Phi}$ are not influenced by the trait-evolutionary model and vice versa. The sections below focus on the conditional posterior $p(\mathbf{\Psi} \mid \mathscr{T}, \mathbf{Y})$.

3.2. Discrete character integration

Many processes of interest can be modeled as the evolution of discrete traits on the tree (Ronquist 2004). Perhaps the most common discrete outcome of interest is location in phylogeographic models (Sanmartín et al. 2008, Comas et al. 2013, Lemey et al. 2020). However, other discrete characters of interest include pathogen host species (Ward et al. 2014, Dearlove et al. 2016, Latinne et al. 2020) and ecological habitat (Bryja et al. 2014, Terra-Araujo et al. 2015, Sánchez-Baracaldo et al. 2017). See Baele et al. (2017b), Table 1 for a more thorough list of discrete-trait analyses.

The most common model of discrete-character evolution is essentially the same as the continuous-time Markov model of nucleotide evolution introduced in Section 1.1. The states can be arbitrarily defined to be whatever discrete character is evolving along the tree.

3.2.1. Developments in Markov jump processes.—

Problems of both genetic sequence and discrete trait evolution have motivated much work on Bayesian networks, hidden Markov models, endpoint-conditioned Markov jump processes and Markov reward processes to infer the number of times specific trait changes occur or the length of time a trait is realized along an evolution history. Siepel et al. (2006), for example, analytically derive the probability mass function of the total number of Markov jumps in an endpoint-conditioned continuous-time Markov chain along a graph with arbitrary rate matrix. Similarly, Minin & Suchard (2008a,b) analytically calculate the moments of the number of jumps between each pair of states. Sometimes, expectations are insufficient and simulation is required to answer the question of interest. Hobolth & Stone (2009) provide several approaches for simulating endpoint-conditioned continuous-time Markov chains. Minin & Suchard (2008a) and Hobolth & Jensen (2011) develop computationally efficient, simulation-free methods for calculating the moments of Markov reward processes (e.g. the average amount of time spent in a particular state of a continuous-time Markov chain).

Phylogenetics has also motivated the development of statistical theory related to Lie Markov models (Sumner et al. 2012, Fernández-Sánchez et al. 2015). These models comprise inhomogeneous continuous-time Markov processes whose endpoint can be expressed as the result of a time-homogeneous process (essentially the time-resolved average of the inhomogeneous process). These processes permit the instantaneous rate matrix to vary over time (and along different branches in a phylogeny) and are useful for identifying the root position of a phylogeny without specifying a molecular clock (Hannaford et al. 2020).

3.2.2. Evolutionary covariates and the curse of dimensionality.—Phylogenetic models are certainly not immune from the curse of dimensionality. This phenomenon is particularly acute in phylogeographic models where the number of discrete locations can be quite large. Assuming a continuous-time Markov process along the phylogeny with L discrete states and infinitesimal rate matrix $\mathbf{Q} = \{q_{\ell m}\}$, the number of free parameters in $\mathbf{Q}$ scales $\mathcal{O}\left(L^2\right)$. While there is no theoretical prohibition on inferring more parameters than there are observations, it becomes increasingly difficult to extract meaningful information in these settings.

This challenge is also an opportunity, as one can reduce the size of the parameter space by assuming the $\mathcal{O}\left(L^2\right)$ transition rates are functions of some low-dimensional process parameterized by scientifically relevant covariates. Lemey et al. (2014) and Zhao et al. (2016) develop a generalized linear model (GLM) that assumes the log-transition rates are a linear function of relevant covariates (e.g. pairwise air traffic between two locations, local temperature) with the number of parameters scaling linearly with the number of covariates. To further penalize over-parameterization within the GLM, Lemey et al. (2014) also assume *a priori* that some unspecified number of covariates have no influence on the transition rates as follows. Let $\mathbf{Z} = \{z_{\ell m,i}\}$ be the covariate observations associated with all ordered pairs $\ell, m \in \{1, \ldots, L\}^2$, $\ell \neq m$ and covariates $i = 1, \ldots, R$. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_R)^t$ be a vector of regression coefficients and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_R)^t$ be a vector of indicator variables such that $\log q_{\ell m} = \sum_{i=1}^{R} \delta_i \beta_i z_{\ell m,i}$. Inference of the indicators i can be achieved via Bayesian stochastic search variable selection (Kuo & Mallick 1998, Chipman et al. 2001). To sample efficiently from a posterior with high correlation between regression coefficients $\boldsymbol{\beta}$, Lemey et al. (2014) rely on a Markov chain transition kernel that draws the proposal $\boldsymbol{\beta}^* \sim \mathcal{N}\left(\boldsymbol{\beta}, \alpha \mathbf{Z}^t \mathbf{Z}\right)$, where α is a tunable scaling factor. This kernel accounts for the prior expectation that coefficients associated with correlated covariates will also be correlated. Zhao et al. (2016), as an alternative, develop an HMC sampler for the regression coefficients. These GLM approaches are applicable beyond phylogenetics and facilitate inference of the rate matrix of any discrete-state continuous-time Markov process.

3.2.3. Piece-wise deterministic, non-reversible Markov processes.—Bouchard-Côté et al. (2018) introduce the bouncy particle sampler (BPS) as a non-reversible, rejection-free alternative to reversible Metropolis-Hastings and HMC samplers. While they evaluate the BPS as a way to efficiently sample from the phylogenetic rate matrix $\mathbf{Q}$, it has broad utility beyond statistical phylogenetics. Inspired by the physics literature (Peters & de With

2012), the BPS relies on piece-wise linear trajectories of a particle (the parameters) through a potential field (the negative log-posterior). Bouchard-Côté et al. (2018) generalize this sampler and develop methods to exactly simulate the parameter trajectories. The BPS relies on finding the parameter value along a line that maximizes the posterior density. Bouchard-Côté et al. (2018) use gradient calculations from the HMC sampler of Zhao et al. (2016) to identify these maxima and sample efficiently from a high-dimensional evolutionary rate matrix. See Section 3.3.2 for additional applications of piece-wise deterministic, non-reversible Markov processes.

3.3. Gaussian processes on a tree

While discrete-trait models discussed above are typically based on the same model of molecular sequences introduced in Section 1.1, continuous data integration requires new statistical models. Due to their computational tractability, Gaussian processes form the backbone of most continuous trait analyses. The simplest such model is one where correlated traits evolve according to a P-dimensional multivariate Brownian diffusion (MBD) process (Edwards & Cavalli-Sforza 1964, Felsenstein 1985b). Using the notation of Section 3.1, we have

$$\mathbf{x}_i \mid \mathbf{x}_{\mathrm{pa}(i)} \sim \mathcal{N}\left(\mathbf{x}_{\mathrm{pa}(i)}, t_i \boldsymbol{\Sigma}\right) \text{ and } \mathbf{y}_i = f(\mathbf{x}_i) = \mathbf{x}_i . \quad 7.$$

Marginalizing the latent traits (except the root traits $\mathbf{x}_{2N-1}$) results in the likelihood

$$\mathrm{vec}(\mathbf{Y}) \mid \mathcal{F}, \mathbf{x}_{2N-1}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\mathrm{vec}(\mathbf{1}_N \mathbf{x}_{2N-1}^t), \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi}\right), \quad 8.$$

where $\otimes$ is the Kronecker product and $\boldsymbol{\Psi}$ is a deterministic function of the phylogenetic tree $\mathcal{F}$ capturing the phylogenetically-induced covariance between taxa.

Likelihood-based inference frequently requires repeated evaluation of the likelihood function $p(\mathbf{Y} \mid \mathcal{F}, \mathbf{x}_{2N-1}, \boldsymbol{\Sigma})$, which naively scales $\mathcal{O}\left(N^3 P^3\right)$. Exploiting the Kronecker product to invert the variance reduces this complexity to $\mathcal{O}\left(N^3 + P^3\right)$. As both N and P can be large, even this greatly simplified calculation can be intractable. Freckleton (2012) (based on Felsenstein (1973b)), Pybus et al. (2012) and Ho & Ané (2014) develop strategies for computing this likelihood in $\mathcal{O}\left(NP^2 + P^3\right)$ using approaches conceptually similar to Felsentein's pruning algorithm for computing the sequence-based likelihood (Felsenstein 1973a). The Ho & Ané (2014) approach uses the tree structure to efficiently compute

$$(\mathbf{Y} - \mathbf{1}_N \mathbf{x}_{2N-1}^t)^t \boldsymbol{\Psi}^{-1} (\mathbf{Y} - \mathbf{1}_N \mathbf{x}_{2N-1}^t) \quad 9.$$

in $\mathcal{O}\left(NP^2\right)$ for any matrix $\mathbf{\Psi}$ that satisfies what they dub the 3-point structure. Specifically, any matrix $\mathbf{\Psi}$ has a 3-point structure if for all i, j, k the two smallest covariances of ψ_{ij}, ψ_{ik}, ψ_{jk} are equal to each other. Ho & Ané (2014) generalize this to allow negative covariances in $\mathbf{\Psi}$ under certain conditions. More recently, Bastide et al. (2020) develop an HMC-based approach that can calculate gradients for nearly all relevant parameters in these hierarchical Gaussian models in linear time.

3.3.1. Gaussian processes and Matrix-Normal likelihoods with missing data.

—Unfortunately, the previous methods for computing the likelihood fail with partially missing data. Cybis et al. (2015) address missing data within a tip in these hierarchical Gaussian process models via data augmentation. Let $\mathbf{y}_i^{\text{mis}}$ and $\mathbf{y}_i^{\text{obs}}$ be the missing and observed data, respectively, associated with tip node v_i. Cybis et al. (2015) develop a procedure that can sample from $\mathbf{y}_i^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathcal{F}, \mathbf{\Sigma}$ for $i = 1, \ldots, N$. Each sample requires $\mathcal{O}\left(NP^2\right)$ computations for $\mathcal{O}\left(N^2P^2\right)$ complexity to sample from all N tips.

Bastide et al. (2018), Mitov et al. (2020) and Hassler et al. (2020) develop an alternative approach that analytically integrates out missing observations rather than relying on data augmentation. This approach assumes that

$$\mathbf{y}_i \mid \mathbf{x}_i \sim \mathcal{N}\left(\mathbf{x}_i, \mathbf{R}_i^t \begin{pmatrix} \infty\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{R}_i\right) \tag{10.}$$

where $\mathbf{R}_i$ is a permutation matrix that arranges the ∞ values to correspond to the indices of $\mathbf{y}_i^{\text{mis}}$ and the 0 values to correspond to the indices of $\mathbf{y}_i^{\text{obs}}$. This specification of missingness gives rise to a series of non-standard operations involving square matrices with 0 or ∞ diagonal elements. For example, the special inverse of some arbitrary matrix

$$\left[\mathbf{R}_i^t \begin{pmatrix} \infty\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{R}_i\right]^{-} = \mathbf{R}_i^t \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \infty\mathbf{I} \end{pmatrix} \mathbf{R}_i. \tag{11.}$$

Propagating missing information up the tree via singular precision matrices allows marginal likelihood calculations of the observed data only in $\mathcal{O}\left(NP^3\right)$.

This algorithm applies to a much broader range of statistical models than MBD on a tree and helps solve the longstanding statistical challenge of efficiently calculating multivariate normal likelihoods with missing data. Specifically, it applies to any multivariate normal likelihood with a 3-point structured covariance matrix discussed above (Ho & Ané 2014). This structure is common in hierarchical Gaussian models. While Allen & Tibshirani (2010) and Glanz & Carvalho (2018) use the expectation-maximization algorithm to perform maximum likelihood imputation, the Bastide et al. (2018)/Mitov et al. (2020)/Hassler et al. (2020) approach permits inference relying on only the observed-data likelihood. For

situations where imputation is desired, this approach allows one to sample from the full conditional distribution of all missing observations simultaneously in $\mathcal{O}\left(NP^3\right)$ time as well.

3.3.2. Multivariate probit models and sampling from high-dimensional truncated Gaussian distributions.—Bayesian phylogenetics has also served as the motivation for many novel methods in multivariate probit models. Cybis et al. (2015) develop a phylogenetically informed multivariate probit model with correlations between both traits and taxa. Under this model, the data are a mix of continuous and discrete traits. Underlying all traits is an MBD process on the tree. Here, the mapping $f(\mathbf{x}_i) = (f_1(x_{i1}), \ldots, f_P(x_{iP}))^t$ between the continuous latent traits $\mathbf{x}_i$ and mixed continuous/discrete observed data $\mathbf{y}_i$ is not the simple identity function. For a binary trait j, we have $y_{ij} = f_j(x_{ij}) = 1_{\{x_{ij} > 0\}}$ (see Cybis et al. (2015) for mappings to ordinal or categorical traits). For continuous traits k, the link function remains $f_k(x_{ij}) = x_{ij}$.

Let $\mathbf{x}_i^{\text{obs}}$ be the components of $\mathbf{x}_i$ associated with the continuous phenotypes and let $\mathbf{x}_i^{\text{lat}}$ be the latent components informing the discrete traits. Efficient inference under this model requires data augmentation of $\mathbf{x}_i^{\text{lat}}$ for $i = 1, \ldots, N$. As mentioned in Section 3.3.1, this procedure relies on sampling from $\mathbf{x}_i^{\text{lat}} \mid \mathbf{y}_i, \mathbf{X}_{\setminus i}, \mathscr{T}, \boldsymbol{\Sigma}$ for $i = 1, \ldots, N$, where $\mathbf{X}_{\setminus i} = \{\mathbf{x}_j; j \neq i\}$. This full conditional posterior is a (potentially high-dimensional) truncated Gaussian distribution due to the constraints in the stochastic link function. While Cybis et al. (2015) rely on a multiple-try rejection sampler, this sampler can be prohibitively slow for high-dimensional truncated Gaussian distributions. Zhang et al. (2021), however, employ a novel approach, the BPS (Bouchard-Côté et al. 2018, see Section 3.2.3), to more efficiently sample from this challenging distribution. As noted previously, the BPS requires calculating the gradient of the log-posterior density with respect to the latent parameters $\mathbf{x}_i^{\text{lat}}$ for $i = 1, \ldots, N$, which Zhang et al. (2021) achieve in linear time with a post-order tree traversal similar to that employed by Pybus et al. (2012). This Zhang et al. (2021) sampler essentially bounces off the truncations of the full conditional posterior. As the truncations are defined on a univariate basis, evaluating when these boundary events occur is trivial, and Zhang et al. (2021) observe increases in computational efficiency over rejection sampling approaching two orders of magnitude.

Seeking improvement on the BPS, Zhang et al. (2022) develop a zigzag Hamiltonian Monte Carlo sampler (Nishimura et al. 2020, zigzag-HMC) to further address the challenge of sampling from a high-dimensional truncated Gaussian distribution in the phylogenetic context. Zigzag-HMC differs from traditional HMC as it posits a Laplace momentum which imparts the unusual property that the Hamiltonian trajectory may only have slopes in $\{\pm 1\}^d$ where d is the dimensionality of the parameter space (i.e. the element-wise slopes may be 1 or −1 only). As the velocity restricted to $\{\pm 1\}^d$ only depends on the sign of the momentum, the particle moves with a constant velocity until one momentum component changes its sign, at which point the particle updates its velocity and moves along a new linear trajectory. See Figure 4 for a simple example. For Gaussian distributions, one can analytically simulate the zigzag Hamiltonian dynamics by calculating when these sign changes occur, eliminating the need for an accept/reject step. Zigzag-HMC handles truncations in the same way as the

BPS and it also takes advantage of the linear time log-posterior gradient evaluations. Besides being more efficient than BPS on a truncated Gaussian, zigzag-HMC also enables a joint update of latent parameters and the across-trait correlation, further improving the sampling efficiency. Importantly, this Zhang et al. (2022) method is able to learn the conditional dependence between any two traits in large problems where BPS fails.

3.3.3. Highly structured, high dimensional data and latent factor models.— Up to this point, we have primarily discussed the computational challenges associated with big-N problems. Big-P data sets are increasingly common in phylogenetic problems, and the methods discussed previously scale at best quadratically in P. Bayesian latent factor models (Press & Shigemasu 1989, Lopes & West 2004) are a common approach to reduce both computational and model complexity. These models assume that the P-dimensional observed data $\mathbf{y}_i$ arise from $K < P$ dimensional latent processes $\mathbf{x}_i$. Specifically, $\mathbf{y}_i = f(\mathbf{x}_i) = \mathbf{L}^t\mathbf{x}_i + \epsilon_i$, where $\mathbf{L}$ is a $K \times P$ estimable matrix and $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \text{diag}[\boldsymbol{\sigma}])$. The standard (non-phylogenetic) model assumes the prior distribution $\mathbf{x}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$, but this specification precludes the requisite correlation between the latent factors that the phylogeny induces. As such, Tolkoff et al. (2018) introduce phylogenetic factor analysis, where the $\mathbf{x}_i$ evolve along the phylogenetic tree via MBD. Standard procedures for sampling from the full conditional posterior of the loadings matrix $\mathbf{L}$ require conditioning on the latent traits $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^t$, and Tolkoff et al. (2018) rely on the procedure outlined in Cybis et al. (2015) to sample from $\mathbf{x}_i \mid \mathbf{y}_i, \mathbf{X}_{\setminus i}, \mathcal{F}, \boldsymbol{\sigma}$ for $i = 1, \ldots, N$ with overall complexity $\mathcal{O}\left(N^2PK^2\right)$. Hassler et al. (2021) apply the likelihood calculation and data augmentation algorithms of Hassler et al. (2020) to sample from $\mathbf{X} \mid \mathbf{Y}, \mathcal{F}, \mathbf{L}, \boldsymbol{\sigma}$ in $\mathcal{O}\left(NPK^3\right)$. As K is by design small, the cubic scaling in K is preferable to the quadratic scaling in N.

Hassler et al. (2021) also develop a novel HMC approach to efficiently sample directly from $\mathbf{L} \mid \mathbf{Y}, \mathcal{F}, \boldsymbol{\sigma}$ without conditioning on the latent factors $\mathbf{X}$ that applies to latent factor models generally. Hassler et al. (2021) show that one can calculate the gradient $\nabla_{\mathbf{L}} \log p(\mathbf{L} \mid \mathbf{Y}, \mathcal{F}, \boldsymbol{\sigma})$ required for HMC as a function of the full conditional mean and variance of each $\mathbf{x}_i$, but not the values of $\mathbf{x}_i$ explicitly. In the phylogenetic context, Hassler et al. (2021) use methods previously developed by Bastide et al. (2018) and Fisher et al. (2021) to calculate these gradients in $\mathcal{O}\left(NPK^3\right)$. This approach is easily transferable to non-phylogenetic latent factor models.

3.3.4. Beyond MBD.— While the continuous trait models discussed above rely on MBD, we emphasize work on other models of continuous evolution. The closely related Ornstein–Uhlenbeck process (Uhlenbeck & Ornstein 1930) is a Gaussian process where traits tend to revert to some mean value (i.e. some evolutionary optimum). Recent work has focused on inferring the points along the phylogeny at which these optima change, known as adaptive shifts (Uyeda & Harmon 2014). Bastide et al. (2018) develop efficient likelihood calculations under a special case of this model. Other models include diffusion on a sphere (Bouckaert 2016) and within a latent space arising from a multidimensional scaling (Holbrook et al. 2021) when only pair-wise distances between traits are observed.

3.4. Preferential sampling and bias

Phylogenetic analyses typically study biological populations evolving in the real world and are inherently observational. As such, data ascertainment is an important factor in any phylogenetic study, with preferential sampling possibly biasing results (Karcher et al. 2016). Phylogeographic models that capture spatiotemporal evolution are particularly susceptible to non-uniform sampling across both space and time (Guindon & De Maio 2021, Kalkauskas et al. 2021). In infectious disease phylogeography, data ascertainment typically requires sequencing the viral genome associated with an individual infection. Unsurprisingly, there are numerous disparities that lead to preferential sampling across both time and space. Both testing and sequencing can be expensive, and resource-rich regions tend to sequence a higher proportion of actual infections (Brito et al. 2021). In the extreme case there may be no sequences available from a location with high levels of known transmission. In addition to sub-sampling to create more representative data sets, researchers have developed several strategies to address bias induced by preferential sampling.

3.4.1. Directly modeling ascertainment.—The coalescent tree priors mentioned in Section 1.2 enable inference of (possibly time-varying) effective population size (EPS). Unsurprisingly, estimates of time-varying EPS are particularly sensitive to preferential sampling in time. While standard models (often inappropriately) assume that sequence ascertainment does not depend on EPS, Karcher et al. (2016) explicitly model ascertainment as an inhomogeneous Poisson process with intensity a function of EPS. They demonstrate via simulation that this approach reduces bias in EPS estimates when sequence ascertainment is proportional to EPS, a common scenario in epidemiological studies.

3.4.2. Sequence-free observations.—When the spatiotemporal distribution of an epidemic can be estimated *a priori*, one can partially correct for preferential sampling by introducing sequence-free samples into the phylogenetic trait reconstruction. Up to this point we have taken for granted that all tip nodes in the phylogeny correspond to an associated molecular sequence as the sequences are the primary source of information for inferring the phylogeny itself. As there are situations where one has access to information about the spatiotemporal distribution of an epidemic (e.g. regional case counts) but relatively few sequences from certain locations, Lemey et al. (2020) and Kalkauskas et al. (2021) propose introducing sequence-free nodes to the phylogenetic tree and demonstrate that this approach can reduce bias induced by extremely biased sampling. Of course, this approach requires prior knowledge of the true spatiotemporal distribution of the process of interest.

3.4.3. Structured coalescent.—An alternative model of discrete phylogeographic migration is the structured coalescent (Notohara 1990), which posits a backward-in-time process where lineages converge and migrate between sub-populations. Where the previously-discussed discrete-trait model assumes the tree is *a priori* independent of the location data, the structured coalescent explicitly models dependence of the tree on the locations, which can reduce bias in both ancestral state reconstructions and rates of migration between locations. As the population demographics are explicit model parameters, they can in turn be informed by other sources of data, further avoiding some biases introduced by preferential sampling of individuals in some states (De Maio et al. 2015).

The primary challenge to inference under these structured coalescent models is that there is no analog to Felsenstein's pruning algorithm (Felsenstein 1973a, 1981, see Section 1.1) that analytically integrates out the migration events. As such, inference under these models requires numerically marginalizing the migration history, typically via MCMC (Vaughan et al. 2014).

De Maio et al. (2015) develop an approximation to the standard structured coalescent model that does allow analytic integration of the migration histories, avoiding laborious numerical integration. Volz (2012) and Müller et al. (2017) also develop efficient numerical approximations of the structured coalescent likelihood. Existing implementations of structured coalescent models, however, still compare poorly computationally with the simpler discrete trait models and are intractable for large-scale problems. Improving computational efficiency in these models is an active area of research.

4. Case Study

Phylogenetics has increasingly played a role in studying viral epidemic dynamics, sometimes in real time (Dellicour et al. 2021, Hodcroft et al. 2021). Researchers can integrate information about the spatiotemporal spread of a virus into phylogenetic models to identify an epidemic's origin (Plantier et al. 2009, Liu et al. 2013, Worobey et al. 2016) and transmission dynamics (Ehichioya et al. 2011, Dudas et al. 2017, Du Plessis et al. 2021). In these phylodynamic analyses, the sampling time and location of a genetic sequence are critical data that allow researchers to reconstruct how a virus spreads through populations.

Here, we consider a case study arising out of the paper by Lemey et al. (2020) on early SARS-CoV-2 international transmission. In addition to viral genetic sequences, sample dates and sample locations, Lemey et al. (2020) incorporate information on individual travel history, global air traffic patterns, local outbreak intensity and within-host infection dynamics. The authors seek to identify the paths along which SARS-CoV-2 traveled as it escaped Hubei province, China, and spread globally. As discussed in Section 3.4, phylogeographic analyses are susceptible to ascertainment bias, which is often unavoidable as viral transmission does not respect administrative boundaries with consistent sequencing and reporting. To address this challenge, Lemey et al. (2020) integrate both individual-level travel history and location-specific estimated case counts into their phylogeographic analysis.

Lemey et al. (2020) collect 282 early SARS-CoV-2 sequences from around the world. Roughly 20% of these sequences were associated with recorded international travel. As they consider 44 discrete locations, they parameterize the transition rate matrix via a GLM with pairwise air traffic connectivity and geographic distance as covariates (see Section 3.2.2). To incorporate travel history, they introduce additional degree-2 internal nodes (i.e. nodes with a single parent and single child) into the phylogeny and assign the travel origins to those nodes. The dates of these nodes are fixed to the travel dates (when known) or inferred assuming a prior informed by the SARS-CoV-2 incubation time. The travel destinations remain assigned to tip nodes. Finally, Lemey et al. (2020) incorporate sequence-free observations from under-sampled locations such as Italy and Iran.

Ultimately, incorporating these various sources of information into the discrete trait phylogeographic model resulted in more plausible transmission patterns and a statistical model with greater out-of-sample predictive performance (see Figure 5). The Bayesian approach allows seamless incorporation of prior knowledge in 1) SARS-CoV-2 case counts informing the locations and dates of sequence-free tip nodes and 2) SARS-CoV-2 within-host dynamics informing the prior on the time between the origin and destination nodes associated with specific travelers. These approaches also permitted accommodation of uncertainty in the phylogenetic tree itself, as the phylogenetic tree was inferred simultaneously with all transmission dynamics via MCMC simulation.

5. Discussion

Phylogenetics has motivated numerous theoretical, methodological, and computational advances in the statistics of Bayesian networks, continuous-time Markov processes and Gaussian processes. The challenges of dealing with complex, hierarchical statistical models with combined continuous/discrete parameter spaces continue to spur creative statistical innovations. Many of the topics discussed are active areas of research.

The Bayesian approach is particularly useful in phylogenetics as the phylogeny itself is frequently a nuisance parameter. Analyses that condition on a single phylogeny do not properly account for the often high degree of uncertainty in the phylogenetic estimates. Numerically marginalizing over the phylogeny via MCMC or other approaches discussed in Section 2 conveniently addresses this uncertainty. Similarly, the Bayesian approach offers a intuitive way to account for uncertainty in the phylogeny. Beyond properly measuring uncertainty, there are cases where we do indeed have prior information about relevant parameter values such as the root date (e.g. the temporal origin of a pandemic) or branch lengths (e.g. rapidly growing populations tend to have shorter branch lengths near the root).

Despite the many advances, there are persistent challenges in both inferring the tree itself and data integration. The SARS-CoV-2 pandemic greatly accelerated previous gains in epidemic genomic surveillance. Bayesian methods are typically limited to several thousand taxa and currently require down-sampling when analyzing some pandemic-scale data sets. Recent work has focused on computationally efficient implementations of simpler models (https://beast.community/thorney_beast) or approximate likelihoods (De Maio et al. 2022). Additionally, as discussed in Section 3.4, common phylogeographic models exhibit a trade-off between computational efficiency and robustness to sampling bias.

Finally, while we focus here on the statistical implications related to data integration in Bayesian phylogenetics, we direct the reader to Baele et al. (2017b) for a thorough discussion of data integration from a more biological perspective with more specific examples.

Acknowledgments

This work was supported through US National Institutes of Health grants HG006139, AI153044, AI154824 and AI162611; the European Research Council under the European Union's Horizon 2020 grant agreement no. 725422-ReservoirDOCS; and Wellcome Trust project 206298/Z/17/Z (Artic Network). GB acknowledges support from the Internal Funds KU Leuven (Grant No. C14/18/094) and the Research Foundation - Flanders ("Fonds voor

Wetenschappelijk Onderzoek – Vlaanderen," G0E1420N, G098321N). PL acknowledges support by the Research Foundation – Flanders ("Fonds voor Wetenschappelijk Onderzoek – Vlaanderen," G0D5117N and G051322N).

LITERATURE CITED

Alfaro ME, Zoller S, Lutzoni F. 2003. Bayes or bootstrap? a simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Molecular Biology and Evolution 20(2):255–266 [PubMed: 12598693]

Allen GI, Tibshirani R. 2010. Transposable regularized covariance models with an application to missing data imputation. The Annals of Applied Statistics 4(2):764 [PubMed: 26877823]

Ayres DL, Cummings MP, Baele G, Darling AE, Lewis PO, et al. 2019. BEAGLE 3: improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. Systematic Biology 68(6):1052–1061 [PubMed: 31034053]

Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, et al. 2012. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Systematic Biology 61(1):170–173 [PubMed: 21963610]

Baele G, Lemey P, Rambaut A, Suchard MA. 2017a. Adaptive MCMC in Bayesian phylogenetics: an application to analyzing partitioned data in BEAST. Bioinformatics 33(12):1798–1805 [PubMed: 28200071]

Baele G, Suchard MA, Rambaut A, Lemey P. 2017b. Emerging concepts of data integration in pathogen phylodynamics. Systematic Biology 66(1):e47–e65 [PubMed: 28173504]

Baldauf SL. 2003. Phylogeny for the faint of heart: a tutorial. Trends in Genetics 19(6):345–351 [PubMed: 12801728]

Barido-Sottani J, Vaughan TG, Stadler T. 2020. A multitype birth–death model for Bayesian inference of lineage-specific birth and death rates. Systematic Biology 69(5):973–986 [PubMed: 32105322]

Bastide P, Ané C, Robin S, Mariadassou M. 2018. Inference of adaptive shifts for multivariate correlated traits. Systematic Biology 67(4):662–680 [PubMed: 29385556]

Bastide P, Ho LST, Baele G, Lemey P, Suchard MA. 2020. Efficient Bayesian inference of general Gaussian models on large phylogenetic trees. arXiv preprint arXiv:2003.10336

Berry V, Gascuel O. 1996. On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. Molecular Biology and Evolution 13(7):999–1011

Bouchard-Côté A, Sankararaman S, Jordan MI. 2012. Phylogenetic inference via sequential Monte Carlo. Systematic Biology 61(4):579–593 [PubMed: 22223445]

Bouchard-Côté A, Vollmer SJ, Doucet A. 2018. The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. Journal of the American Statistical Association 113(522):855–867

Bouckaert R 2016. Phylogeography by diffusion on a sphere: whole world phylogeography. PeerJ 4:e2406 [PubMed: 27651992]

Bradley BJ. 2008. Reconstructing phylogenies and phenotypes: a molecular view of human evolution. Journal of Anatomy 212(4):337–353 [PubMed: 18380860]

Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, et al. 2021. Global disparities in SARS-CoV-2 genomic surveillance. medRxiv

Bryja J, Mikula O, Šumbera R, Meheretu Y, Aghová T, et al. 2014. Pan-African phylogeny of Mus (subgenus Nannomys) reveals one of the most successful mammal radiations in Africa. BMC Evolutionary Biology 14(1):1–20

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, et al. 2017. Stan: A probabilistic programming language. Journal of Statistical Software 76(1)

Chipman H, George EI, McCulloch RE, Clyde M, Foster DP, Stine RA. 2001. The practical implementation of Bayesian model selection. IMS Lecture Notes-Monograph Series :65–134

Comas I, Coscolla M, Luo T, Borrell S, Holt KE, et al. 2013. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nature Genetics 45(10):1176–1182 [PubMed: 23995134]

Cybis GB, Sinsheimer JS, Bedford T, Mather AE, Lemey P, Suchard MA. 2015. Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. The Annals of Applied Statistics 9(2):969 [PubMed: 27053974]

Dang T, Kishino H. 2019. Stochastic variational inference for Bayesian phylogenetics: a case of CAT model. Molecular Biology and Evolution 36(4):825–833 [PubMed: 30715448]

De Maio N, Kalaghatgi P, Turakhia Y, Corbett-Detig R, Minh BQ, Goldman N. 2022. Maximum likelihood pandemic-scale phylogenetics. bioRxiv

De Maio N, Wu CH, O'Reilly KM, Wilson D. 2015. New routes to phylogeography: a Bayesian structured coalescent approximation. PLoS Genetics 11(8):e1005421 [PubMed: 26267488]

Dearlove BL, Cody AJ, Pascoe B, Méric G, Wilson DJ, Sheppard SK. 2016. Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. The ISME Journal 10(3):721–729 [PubMed: 26305157]

Dellicour S, Baele G, Dudas G, Faria NR, Pybus OG, et al. 2018. Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. Nature Communications 9(1):1–9

Dellicour S, Durkin K, Hong SL, Vanmechelen B, Martí-Carreras J, et al. 2021. A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of SARS-CoV-2 lineages. Molecular Biology and Evolution 38(4):1608–1613 [PubMed: 33316043]

Dinh V, Bilge A, Zhang C, Matsen IV FA. 2017. Probabilistic path Hamiltonian Monte Carlo, In International Conference on Machine Learning, pp. 1009–1018, PMLR

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology 4(5):e88 [PubMed: 16683862]

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biology 8(1):1–12 [PubMed: 20051105]

Du Plessis L, McCrone JT, Zarebski AE, Hill V, Ruis C, et al. 2021. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science 371(6530):708–712 [PubMed: 33419936]

Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, et al. 2017. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature 544(7650):309–315 [PubMed: 28405027]

Edwards A, Cavalli-Sforza L. 1964. Reconstruction of evolutionary trees. In Phenetic and Phylogenetic Classification, eds. Heywood VH, McNeill J. The Systematics Association, 67–76

Efron B, Halloran E, Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. Proceedings of the National Academy of Sciences 93(14):7085–7090

Ehichioya DU, Hass M, Becker-Ziaja B, Ehimuan J, Asogun DA, et al. 2011. Current molecular epidemiology of lassa virus in Nigeria. Journal of Clinical Microbiology 49(3):1157–1161 [PubMed: 21191050]

Fabreti LG, Höhna S. 2021. Convergence assessment for Bayesian phylogenetic analysis using MCMC simulation. bioRxiv

Faulkner JR, Magee AF, Shapiro B, Minin VN. 2020. Horseshoe-based Bayesian nonparametric estimation of effective population size trajectories. Biometrics 76(3):677–690 [PubMed: 32277713]

Felsenstein J 1973a. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Systematic Biology 22(3):240–249

Felsenstein J 1973b. Maximum-likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics 25(5):471 [PubMed: 4741844]

Felsenstein J 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17(6):368–376 [PubMed: 7288891]

Felsenstein J 1985a. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39(4):783–791 [PubMed: 28561359]

Felsenstein J 1985b. Phylogenies and the comparative method. The American Naturalist 125(1):1–15

Felsenstein J 2004. Inferring phylogenies, vol. 2. Sinauer Associates Sunderland, MA

Felsenstein J, Kishino H. 1993. Is there something wrong with the bootstrap on phylogenies? a reply to Hillis and Bull. Systematic Biology 42(2):193–200

Fernández-Sánchez J, Sumner JG, Jarvis PD, Woodhams MD. 2015. Lie Markov models with purine/pyrimidine symmetry. Journal of Mathematical Biology 70(4):855–891 [PubMed: 24723068]

Fisher AA, Ji X, Zhang Z, Lemey P, Suchard MA. 2021. Relaxed random walks at scale. Systematic Biology 70(2):258–267 [PubMed: 32687171]

Flouri T, Izquierdo-Carrasco F, Darriba D, Aberer AJ, Nguyen LT, et al. 2015. The phylogenetic likelihood library. Systematic Biology 64(2):356–362 [PubMed: 25358969]

Fourment M, Claywell BC, Dinh V, McCoy C, Matsen IV FA, Darling AE. 2018. Effective online Bayesian phylogenetics via sequential Monte Carlo with guided proposals. Systematic Biology 67(3):490–502 [PubMed: 29186587]

Fourment M, Darling AE. 2019. Evaluating probabilistic programming and fast variational Bayesian inference in phylogenetics. PeerJ 7:e8272 [PubMed: 31976168]

Fourment M, Magee AF, Whidden C, Bilge A, Matsen IV FA, Minin VN. 2020. 19 dubious ways to compute the marginal likelihood of a phylogenetic tree topology. Systematic Biology 69(2):209–220 [PubMed: 31504998]

Freckleton RP. 2012. Fast likelihood calculations for comparative analyses. Methods in Ecology and Evolution 3(5):940–947

Gaya E, Redelings BD, Navarro-Rosinés P, Llimona X, De Cáceres M, Lutzoni F. 2011. Align or not to align? resolving species complexes within the Caloplaca saxicola group as a case study. Mycologia 103(2):361–378 [PubMed: 21139031]

Gelfand AE. 2000. Gibbs sampling. Journal of the American Statistical Association 95(452):1300–1304

Glanz H, Carvalho L. 2018. An expectation–maximization algorithm for the matrix normal distribution with an application in remote sensing. Journal of Multivariate Analysis 167:31–48

Guindon S, De Maio N. 2021. Accounting for spatial sampling patterns in Bayesian phylogeography. Proceedings of the National Academy of Sciences 118(52)

Haario H, Saksman E, Tamminen J. 2001. An adaptive Metropolis algorithm. Bernoulli :223–242

Hannaford NE, Heaps SE, Nye TM, Williams TA, Embley TM. 2020. Incorporating compositional heterogeneity into Lie Markov models for phylogenetic inference. The Annals of Applied Statistics 14(4):1964–1983

Harrington SM, Wishingrad V, Thomson RC. 2021. Properties of Markov chain Monte Carlo performance across many empirical alignments. Molecular Biology and Evolution 38(4):1627–1640 [PubMed: 33185685]

Hassler G, Tolkoff MR, Allen WL, Ho LST, Lemey P, Suchard MA. 2020. Inferring phenotypic trait evolution on large trees with many incomplete measurements. Journal of the American Statistical Association :1–15

Hassler GW, Gallone B, Aristide L, Allen WL, Tolkoff MR, et al. 2021. Principled, practical, flexible, fast: a new approach to phylogenetic factor analysis. arXiv preprint arXiv:2107.01246

Hillis DM, Bull JJ. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology 42(2):182–192

Ho LST, Ané C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Systematic Biology 63(3):397–408 [PubMed: 24500037]

Hobolth A, Jensen JL. 2011. Summary statistics for endpoint-conditioned continuous-time Markov chains. Journal of Applied Probability 48(4):911–924

Hobolth A, Stone EA. 2009. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. The Annals of Applied Statistics 3(3):1204 [PubMed: 20148133]

Hodcroft EB, Zuber M, Nadeau S, Vaughan TG, Crawford KH, et al. 2021. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. Nature 595(7869):707–712 [PubMed: 34098568]

Höhna S, Drummond AJ. 2012. Guided tree topology proposals for Bayesian phylogenetic inference. Systematic Biology 61(1):1–11 [PubMed: 21828081]

Höhna S, Freyman WA, Nolen Z, Huelsenbeck JP, May MR, Moore BR. 2019. A Bayesian approach for estimating branch-specific speciation and extinction rates. bioRxiv :555805

Holbrook AJ, Lemey P, Baele G, Dellicour S, Brockmann D, et al. 2021. Massive parallelization boosts big Bayesian multidimensional scaling. Journal of Computational and Graphical Statistics 30(1):11–24 [PubMed: 34168419]

Huelsenbeck JP, Rannala B, Masly JP. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. Science 288(5475):2349–2350 [PubMed: 10875916]

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17(8):754–755 [PubMed: 11524383]

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294(5550):2310–2314 [PubMed: 11743192]

Izquierdo-Carrasco F, Alachiotis N, Berger S, Flouri T, Pissis SP, Stamatakis A. 2013. A generic vectorization scheme and a GPU kernel for the phylogenetic likelihood library, In 2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum, pp. 530–538, IEEE

Ji X, Zhang Z, Holbrook A, Nishimura A, Baele G, et al. 2020. Gradients do grow on trees: a linear-time O(N)-dimensional gradient for statistical phylogenetics. Molecular Biology and Evolution 37(10):3047–3060 [PubMed: 32458974]

Kalkauskas A, Perron U, Sun Y, Goldman N, Baele G, et al. 2021. Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. PLoS Computational Biology 17(1):e1008561 [PubMed: 33406072]

Karcher MD, Palacios JA, Bedford T, Suchard MA, Minin VN. 2016. Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. PLoS Computational Biology 12(3):e1004789 [PubMed: 26938243]

Ki C, Terhorst J. 2022. Variational phylodynamic inference using pandemic-scale data. BioRxiv

Kingman JFC. 1982. The coalescent. Stochastic Processes and their Applications 13(3):235–248

Kschischang FR, Frey BJ, Loeliger HA. 2001. Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory 47(2):498–519

Kuo L, Mallick B. 1998. Variable selection for regression models. Sankhya¯: The Indian Journal of Statistics, Series B :65–81

Lakner C, Van Der Mark P, Huelsenbeck JP, Larget B, Ronquist F. 2008. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. Systematic Biology 57(1):86–103 [PubMed: 18278678]

Lanfear R, Hua X, Warren DL. 2016. Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. Genome Biology and Evolution 8(8):2319–2332 [PubMed: 27435794]

Larget B, Simon DL. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution 16(6):750–759

Latinne A, Hu B, Olival KJ, Zhu G, Zhang L, et al. 2020. Origin and cross-species transmission of bat coronaviruses in China. Nature Communications 11(1):1–15

Lemey P, Hong SL, Hill V, Baele G, Poletto C, et al. 2020. Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. Nature Communications 11(1):1–14

Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, et al. 2014. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathogens 10(2):e1003932 [PubMed: 24586153]

Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009a. Bayesian phylogeography finds its roots. PLoS Computational Biology 5(9):e1000520 [PubMed: 19779555]

Lemey P, Rambaut A, Welch JJ, Suchard MA. 2010. Phylogeography takes a relaxed random walk in continuous space and time. Molecular Biology and Evolution 27(8):1877–1885 [PubMed: 20203288]

Lemey P, Ruktanonchai N, Hong SL, Colizza V, Poletto C, et al. 2021. Untangling introductions and persistence in COVID-19 resurgence in europe. Nature 595(7869):713–717 [PubMed: 34192736]

Lemey P, Salemi M, Vandamme AM. 2009b. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge University Press

Li S, Pearl DK, Doss H. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. Journal of the American Statistical Association 95(450):493–508

Liu D, Shi W, Shi Y, Wang D, Xiao H, et al. 2013. Origin and diversity of novel avian influenza A H7N9 viruses causing human infection: phylogenetic, structural, and coalescent analyses. The Lancet 381(9881):1926–1932

Lopes HF, West M. 2004. Bayesian model assessment in factor analysis. Statistica Sinica :41–67

MacPherson A, Louca S, McLaughlin A, Joy JB, Pennell MW. 2022. Unifying phylogenetic birth–death models in epidemiology and macroevolution. Systematic Biology 71(1):172–189

Magee AF, Karcher MD, Matsen IV FA, Minin VN. 2021. How trustworthy is your tree? Bayesian phylogenetic effective sample size through the lens of Monte Carlo error. arXiv preprint arXiv:2109.07629

Mau B, Newton MA, Larget B. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55(1):1–12 [PubMed: 11318142]

May MR, Höhna S, Moore BR. 2016. A Bayesian approach for detecting the impact of mass-extinction events on molecular phylogenies when rates of lineage diversification may vary. Methods in Ecology and Evolution 7(8):947–959

Meyer X 2021. Adaptive tree proposals for Bayesian phylogenetic inference. Systematic Biology 70(5):1015–1032 [PubMed: 33515248]

Minin VN, Bloomquist EW, Suchard MA. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Molecular Biology and Evolution 25(7):1459–1471 [PubMed: 18408232]

Minin VN, Suchard MA. 2008a. Counting labeled transitions in continuous-time Markov models of evolution. Journal of Mathematical Biology 56(3):391–412 [PubMed: 17874105]

Minin VN, Suchard MA. 2008b. Fast, accurate and simulation-free stochastic mapping. Philosophical Transactions of the Royal Society B: Biological Sciences 363(1512):3985–3995

Mitov V, Bartoszek K, Asimomitis G, Stadler T. 2020. Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts. Theoretical Population Biology 131:66–78 [PubMed: 31805292]

Mohasel Afshar H, Domke J. 2015. Reflection, refraction, and Hamiltonian Monte Carlo. Advances in Neural Information Processing Systems 28

Moretti AK, Zhang L, Naesseth CA, Venner H, Blei D, Pe'er I. 2021. Variational combinatorial sequential Monte Carlo methods for Bayesian phylogenetic inference. arXiv preprint arXiv:2106.00075

Müller NF, Rasmussen DA, Stadler T. 2017. The structured coalescent and its approximations. Molecular Biology and Evolution 34(11):2970–2981 [PubMed: 28666382]

Nascimento FF, Reis Md, Yang Z. 2017. A biologist's guide to Bayesian phylogenetic analysis. Nature Ecology & Evolution 1(10):1446–1454 [PubMed: 28983516]

Nee S, May RM, Harvey PH. 1994. The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 344(1309):305–311 [PubMed: 7938201]

Nishimura A, Dunson DB, Lu J. 2020. Discontinuous Hamiltonian Monte Carlo for discrete parameters and discontinuous likelihoods. Biometrika 107(2):365–380

Notohara M 1990. The coalescent and the genealogical process in geographically structured population. Journal of Mathematical Biology 29(1):59–75 [PubMed: 2277236]

Pakman A, Paninski L. 2013. Auxiliary-variable exact Hamiltonian Monte Carlo samplers for binary distributions. Advances in Neural Information Processing Systems 26

Pearl J 1982. Reverend Bayes on inference engines: A distributed hierarchical approach, In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 133–136, Association for the Advancement of Artificial Intelligence

Peters EAJF, de With G. 2012. Rejection-free Monte Carlo sampling for general potentials. Phys. Rev. E 85(2):026703

Plantier JC, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, et al. 2009. A new human immunodeficiency virus derived from gorillas. Nature Medicine 15(8):871–872

Press SJ, Shigemasu K. 1989. Bayesian inference in factor analysis. In Contributions to Probability and Statistics. Springer, 271–287

Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, et al. 2012. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. Proceedings of the National Academy of Sciences 109(37):15066–15071

Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of Molecular Evolution 43(3):304–311 [PubMed: 8703097]

Roberts GO, Rosenthal JS. 2009. Examples of adaptive MCMC. Journal of Computational and Graphical Statistics 18(2):349–367

Ronquist F 2004. Bayesian inference of character evolution. Trends in Ecology & Evolution 19(9):475–481 [PubMed: 16701310]

Sánchez-Baracaldo P, Raven JA, Pisani D, Knoll AH. 2017. Early photosynthetic eukaryotes inhabited low-salinity habitats. Proceedings of the National Academy of Sciences 114(37):E7737–E7745

Sanmartín I, Van Der Mark P, Ronquist F. 2008. Inferring dispersal: a Bayesian approach to phylogeny-based island biogeography, with special reference to the canary islands. Journal of Biogeography 35(3):428–449

Siepel A, Pollard KS, Haussler D. 2006. New methods for detecting lineage-specific selection, In Annual International Conference on Research in Computational Molecular Biology, pp. 190–205, Springer

Sinsheimer JS, Lake JA, Little RJ. 1996. Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. Biometrics :193–210 [PubMed: 8934592]

Stadler T 2010. Sampling-through-time in birth–death trees. Journal of Theoretical Biology 267(3):396–404 [PubMed: 20851708]

Stadler T 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proceedings of the National Academy of Sciences 108(15):6187–6192

Stadler T, Kouyos R, von Wyl V, Yerly S, Böni J, et al. 2012. Estimating the basic reproductive number from viral sequence data. Molecular Biology and Evolution 29(1):347–357 [PubMed: 21890480]

Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. 2013. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences 110(1):228–233

Strimmer K, Pybus OG. 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. Molecular Biology and Evolution 18(12):2298–2305 [PubMed: 11719579]

Suchard MA, Rambaut A. 2009. Many-core algorithms for statistical phylogenetics. Bioinformatics 25(11):1370–1376 [PubMed: 19369496]

Suchard MA, Weiss RE, Sinsheimer JS. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Molecular Biology and Evolution 18(6):1001–1013 [PubMed: 11371589]

Sullivan J, Joyce P. 2005. Model selection in phylogenetics. Annual Review of Ecology, Evolution, and Systematics 36(1):445–466

Sumner JG, Fernández-Sánchez J, Jarvis PD. 2012. Lie Markov models. Journal of Theoretical Biology 298:16–31 [PubMed: 22212913]

Teh Y, Daume H III, Roy DM. 2007. Bayesian agglomerative clustering with coalescents, In Advances in Neural Information Processing Systems, eds. Platt J, Koller D, Singer Y, Roweis S, vol. 20. Curran Associates, Inc.

Terra-Araujo MH, de Faria AD, Vicentini A, Nylinder S, Swenson U. 2015. Species tree phylogeny and biogeography of the neotropical genus Pradosia (Sapotaceae, Chrysophylloideae). Molecular Phylogenetics and Evolution 87:1–13 [PubMed: 25797923]

Thompson EA, Thompson K, Thompson E, et al. 1975. Human evolutionary trees. CUP Archive

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Molecular Biology and Evolution 15(12):1647–1657 [PubMed: 9866200]

Tolkoff MR, Alfaro ME, Baele G, Lemey P, Suchard MA. 2018. Phylogenetic factor analysis. Systematic Biology 67(3):384–399 [PubMed: 28950376]

Uhlenbeck GE, Ornstein LS. 1930. On the theory of the Brownian motion. Physical Review 36(5):823

Uyeda JC, Harmon LJ. 2014. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. Systematic Biology 63(6):902–918 [PubMed: 25077513]

Vaughan TG, Kühnert D, Popinga A, Welch D, Drummond AJ. 2014. Efficient Bayesian inference under the structured coalescent. Bioinformatics 30(16):2272–2279 [PubMed: 24753484]

Volz EM. 2012. Complex population dynamics and the coalescent under neutrality. Genetics 190(1):187–201 [PubMed: 22042576]

Wang L, Bouchard-Côté A, Doucet A. 2015. Bayesian phylogenetic inference using a combinatorial sequential Monte Carlo method. Journal of the American Statistical Association 110(512):1362–1374

Wang L, Wang S, Bouchard-Côté A. 2020. An annealed sequential Monte Carlo method for Bayesian phylogenetics. Systematic Biology 69(1):155–183 [PubMed: 31173141]

Ward M, Gibbons C, McAdam P, Van Bunnik B, Girvan E, et al. 2014. Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of Staphylococcus aureus clonal complex 398. Applied and Environmental Microbiology 80(23):7275–7282 [PubMed: 25239891]

Whidden C, Matsen IV FA. 2015. Quantifying MCMC exploration of phylogenetic tree space. Systematic Biology 64(3):472–491 [PubMed: 25631175]

Worobey M, Watts TD, McKay RA, Suchard MA, Granade T, et al. 2016. 1970s and 'patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. Nature 539(7627):98–101 [PubMed: 27783600]

Yang Z, Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Molecular Biology and Evolution 14(7):717–724 [PubMed: 9214744]

Zhang C 2020. Improved variational Bayesian phylogenetic inference with normalizing flows. Advances in Neural Information Processing Systems 33:18760–18771

Zhang C, Huelsenbeck JP, Ronquist F. 2020. Using parsimony-guided tree proposals to accelerate convergence in Bayesian phylogenetic inference. Systematic Biology 69(5):1016–1032 [PubMed: 31985810]

Zhang C, Matsen IV FA. 2018a. Generalizing tree probability estimation via Bayesian networks. Advances in Neural Information Processing Systems 31

Zhang C, Matsen IV FA. 2018b. Variational Bayesian phylogenetic inference, In International Conference on Learning Representations

Zhang Z, Nishimura A, Bastide P, Ji X, Payne RP, et al. 2021. Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models. The Annals of Applied Statistics 15(1):230–251

Zhang Z, Nishimura A, Ji X, Lemey P, Suchard MA. 2022. Hamiltonian zigzag speeds up large-scale learning of direct effects among mixed-type biological traits. arXiv preprint arXiv:2201.07291

Zhao T, Wang Z, Cumberworth A, Gsponer J, de Freitas N, Bouchard-Côté A. 2016. Bayesian analysis of continuous time Markov chains with application to phylogenetic modeling. Bayesian Analysis 11(4):1203–1237

Zuckerkandl E, Pauling L. 1962. Molecular disease, evolution and genetic heterogeneity. Academic Press

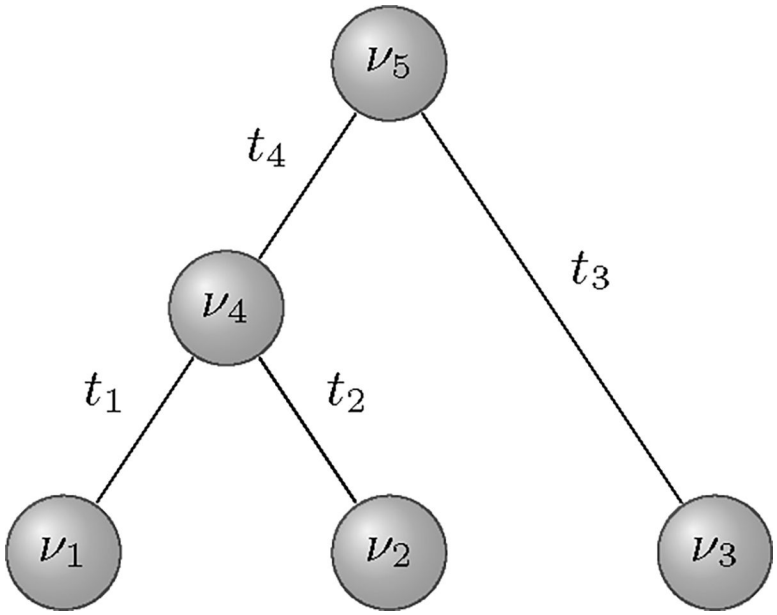

Figure 1:

Simple phylogeny with $N = 3$ degree-one tip nodes $\nu_1, \ldots, \nu_3$, $N - 2 = 1$ degree-three internal node ν_4 and degree-two root node ν_5. The edge connecting each node ν_i to its parent $\nu_{\text{pa}(i)}$ has length t_i. The phylogeny is a directed acyclic graph. It is directed in that there is a parent/child relationship between all nodes connected by an edge, and it is acyclic in that there are no cycles or loops in the graph. Each node has exactly one parent (except for the root which has none).

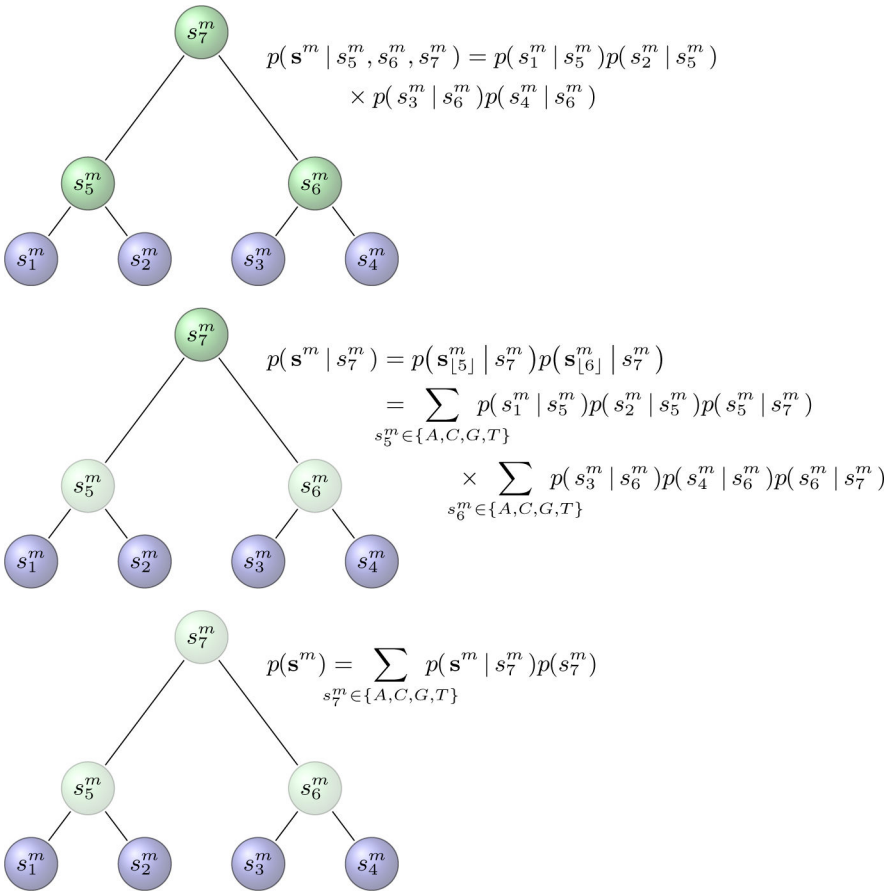

Figure 2:
Example of how Felsenstein's pruning algorithm marginalizes over the ancestral sequences. Tip nodes in blue represent observed sequence data, while green internal nodes represent latent ancestral sequences. Pale nodes have been marginalized. We do not explicitly condition on the tree $\mathcal{F}$ for notational simplicity.

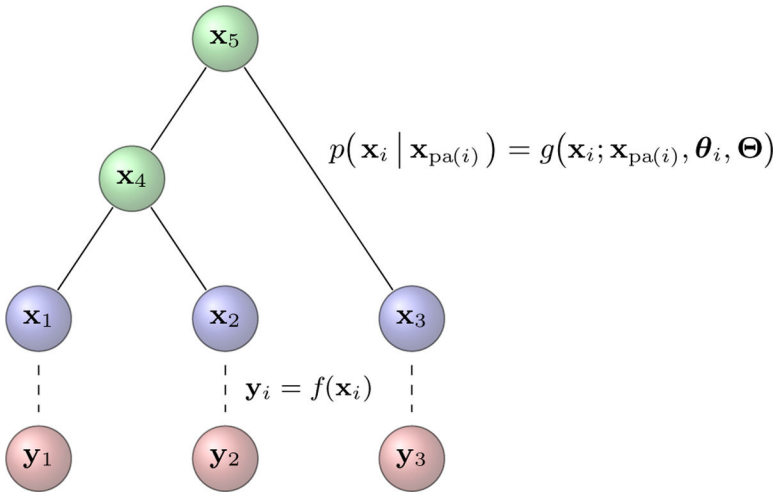

Figure 3:

Schematic of a generalized phylogenetic model. The data $\mathbf{y}_1, \ldots, \mathbf{y}_N$ (red nodes) are assumed to have arisen from the latent traits $\mathbf{x}_1, \ldots, \mathbf{x}_N$ (blue nodes) at the respective tips via the possibly stochastic link function $f(.)$. The latent tip traits $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and latent internal traits $\mathbf{x}_{N+1}, \ldots, \mathbf{x}_{2N-2}$ arise from some evolutionary process on the phylogenetic tree where the traits of each child node $\mathbf{x}_i$ are drawn from a distribution with density $p(\mathbf{x}_i \mid \mathbf{x}_{\mathrm{pa}(i)}) = g(\mathbf{x}_i; \mathbf{x}_{\mathrm{pa}(i)}, \boldsymbol{\theta}_i, \boldsymbol{\Theta})$.

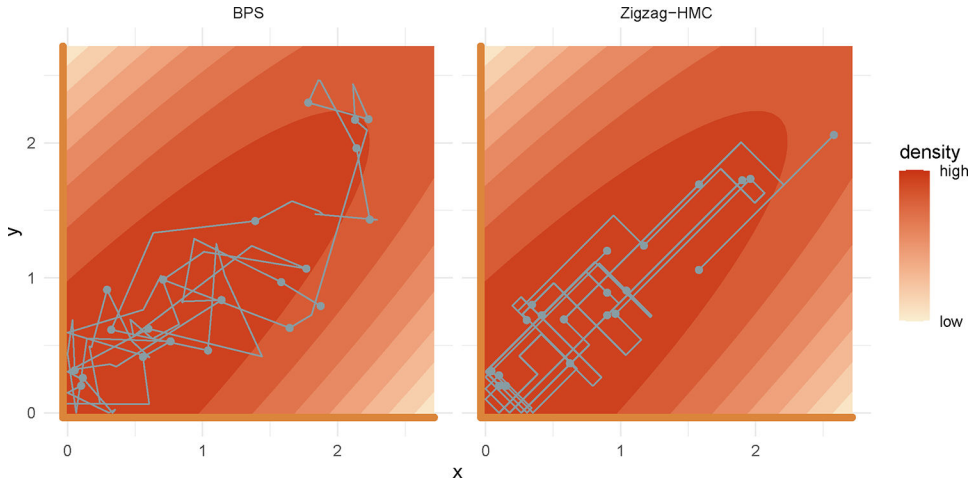

Figure 4:
Sampling from a 2-dimensional truncated Gaussian distribution using both the BPS (left) and zigzag-HMC (right) samplers. Orange lines represent the truncations. Grey lines represent the particle trajectories, while grey dots represent samples from the posterior.

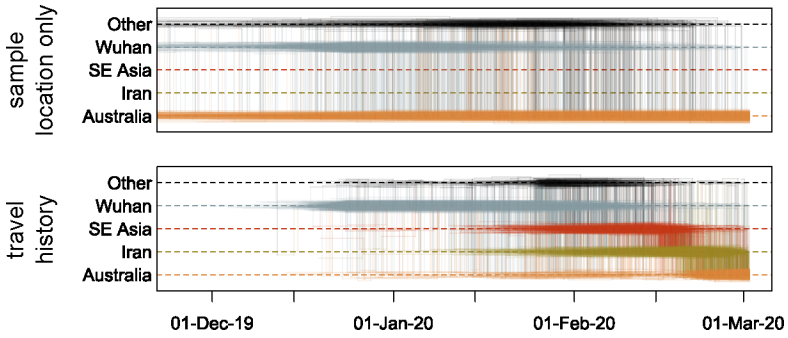

Figure 5:

A toy example of the influence of travel history on discrete trait analyses. Horizontal lines represent persistent lineages within a location, while vertical lines represent transitions between locations in the Markov chain. We inferred a tree with 9 sequences (3 each from Wuhan, Australia, and Europe) where some of the infected individuals sampled in Australia had traveled from Iran or Southeast (SE) Asia. The analysis incorporating travel history captures more information in that the virus is present in all locations and there is less variance in the dates of transition events. This figure was modeled on the tutorial presented in the BEAST documentation. Please note that this is a toy analysis and should not be interpreted as providing insight into the early spread of SARS-CoV-2.