Inferring Phenotypic Trait Evolution on Large Trees With Many Incomplete Measurements

Gabriel Hassler[1],

Max R. Tolkoff[2],

William L. Allen[3],

Lam Si Tung Ho[4],

Philippe Lemey[5],

and Marc A. Suchard[1,2,6]

[1]*Department of Biomathematics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, United States*

[2]*Department of Biostatistics, Jonathan and Karin Fielding School of Public Health, University of California, Los Angeles, United States*

[3]*Department of Biosciences, Swansea University, Swansea, United Kingdom*

[4]*Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada*

[5]*Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium*

[6]*Department of Human Genetics, David Geffen School of Medicine at UCLA, Universtiy of California, Los Angeles, United States*

Corresponding author: Marc A. Suchard, Departments of Biostatistics, Biomathematics, and Human Genetics, University of California, Los Angeles, 695 Charles E. Young Dr., South, Los Angeles, CA 90095-7088, USA; E-mail: msuchard@ucla.edu

Abstract   Comparative biologists are often interested in inferring covariation between multiple biological traits sampled across numerous related taxa. To properly study these relationships, we must control for the shared evolutionary history of the taxa to avoid spurious inference. Existing control techniques almost universally scale poorly as the number of taxa increases. An additional challenge arises as obtaining a full suite of measurements becomes increasingly difficult with increasing taxa. This typically necessitates data imputation or integration that further exacerbates scalability. We propose an inference technique that integrates out missing measurements analytically and scales linearly with the number of taxa by using a post-order traversal algorithm under a multivariate Brownian diffusion (MBD) model to characterize trait evolution. We further exploit this technique to extend the MBD model to account for sampling error or non-heritable residual variance. We test these methods to examine mammalian life history traits, prokaryotic genomic and phenotypic traits, and HIV infection traits. We find computational efficiency increases that top two orders-of-magnitude over current best practices. While we focus on the utility of this algorithm in phylogenetic comparative methods, our approach generalizes to solve long-standing challenges in computing the likelihood for matrix-normal and multivariate normal distributions with missing data at scale.

Keywords: Bayesian inference, matrix-normal, missing data, phylogenetics

1.   INTRODUCTION

Phylogenetic comparative methods explore the relationships between different biological phenotypes across sets of organisms. To properly understand these phenotypic trait relationships, methods must adjust for the shared evolutionary history of the taxa (Felsenstein 1985). Molecular sequences from emerging sequencing technology and high-throughput biological experimentation enable such phylogenetic adjustment for rapidly growing numbers of taxa and increasing numbers of trait measurements. Comparative studies incorporating dense taxonomic sampling create the potential for new research into general patterns in phenotypic evolution, key differences between subgroups and the relationship between phenotypic and genetic evolutionary dynamics. Unfortunately, many phylogenetic comparative methods remain poorly equipped to handle these research questions at scale.

Popular methods often assume an underlying Brownian diffusion process acts along each branch of a phylogenetic tree, such that the traits are multivariate normally distributed. Revell (2012) and Adams (2014), for example, parameterize this distribution in terms of a highly-structured variance-covariance matrix that characterizes the tree and trait covaria-

tion. Computational work to invert this matrix to evaluate the multivariate normal likelihood scales cubically with the number of taxa. This work stands even more troublesome when the phylogenetic tree remains unknown and requires joint inference with the trait process, necessitating repeated inversion. Freckleton (2012), Pybus et al. (2012), and Ho and Ané (2014) all independently develop algorithms that take advantage of the matrix-normal structure of the data under the MBD model to evaluate the likelihood. Using the tree structure, these algorithms then scale linearly with the number of taxa with complete data, but this ideal run-time currently stumbles when trait measurements are missing.

As the number of taxa grows large, measuring a complete suite of traits for all taxa becomes increasingly challenging. Recent solutions to this problem in phylogenetics include those proposed by Tolkoff et al. (2017), who use Markov chain Monte Carlo (MCMC) to numerically integrate via sampling the missing data, and Bastide et al. (2018), who exploit expectation-maximization (EM) to impute missing values and all unobserved ancestral node traits. Both lines of work, however, require iterative manipulation of the likelihood function on a per-taxon basis. Current sampling methods scale at best quadratically with the number of taxa, and likewise the number of ancestral node trait values that require imputation is directly related to the number of taxa. As a consequence, these methods remain computationally prohibitive for large trees.

In this paper, we reformulate evaluation of the data likelihood function under a Brownian diffusion process on a tree such that we achieve the marginalized likelihood of the observed trait measurements only. This innovation arises from thinking about observed tip traits as multivariate normally distributed with infinite precision in their sampling, while missing traits have zero precision, and appropriately propagating these precisions up the tree through dynamic programming involving an unusual matrix pseudo-inverse definition. This pseudo-inverse finds similar use, but independent discovery, in Bastide et al. (2018). Unlike previous approaches, the integration avoids EM iteration making simultaneous inference with the phylogeny practical and enables researchers to analyze all available measurements when inferring the trait relationships. Surprisingly, we can still evaluate the observed-data likelihood in linear time with respect to the number of taxa. The price to be paid is that computation now scales cubically, rather than quadratically, in the number of traits. This remains a small price since the number of taxa is often orders-of-magnitude larger than the number of traits. It is also notable that this method has applications beyond phylogenetic comparative methods and can be used more generally in a special class of matrix-normal and multivariate normal distributions with missing data. This has been a long standing problem in statistics since at least the 1930's (Wilks 1932), with more recent work by Dominici et al. (2000); Cantet et al. (2004); Allen and Tibshirani (2010); and Glanz and Carvalho (2018).

One important limitation to our approach is that it assumes data are missing at random (Little and Rubin 1987) which is inappropriate for many data sets.

We also demonstrate how this framework can be easily extended to incorporate residual variance in the MBD model, which is only one of many possible model extensions. Our strategy of analytically marginalizing the observed data likelihood extends seamlessly to this and other model extensions and allows for efficient inference on these models while maintaining likelihood computations that scale linearly with the number of taxa. These extensions open up lines of inquiry not available in the simple MBD model. In particular, including residual variance in the model enables inference of phylogenetic heritability.

We demonstrate the broad utility of our algorithm to compute the marginalized likelihood through three examples. First, we examine covariation in mammalian life history traits using data on 3690 taxa from the PanTHERIA ecological database (Jones et al. 2009). Second, we use our new efficient algorithm to simultaneously evaluate several theories regarding prokaryotic evolutionary theory. We use data from NCBI Genome and a recent study by Goberna and Verdú (2016), along with matching 16S sequences from the ARB Silva Database (Ludwig et al. 2004), to jointly infer both the phylogenetic tree and evolutionary correlation between several prokaryotic genotypic and phenotypic traits. Finally, we apply our multivariate residual variance model extension to data presented by Blanquart et al. (2017) concerning HIV virulence to evaluate the heritability of HIV viral load and CD4 T-cell decline. We compare the computation speed of our analytical integration method against current best-practice methods and observed increases in speed that top two orders-of-magnitude.

2. PHENOTYPIC DIFFUSION ON TREES

Consider a data-complete collection $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)^t$ where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iP})^t$ of P real-valued phenotypic traits measured across N biological taxa. Relating the taxa stands a known and fixed or unknown and random phylogeny $\mathscr{F}$ that is a bifurcating, directed acyclic graph whose $2N - 1$ vertices originate with a degree-2 root node ν_{2N-1} and terminate with degree-1 tip nodes $(\nu_1, \ldots, \nu_N)$ that correspond to the N taxa. Linking vertices are edge weights or branch lengths $(t_1, \ldots, t_{2N-2})$. Let $\mathbf{X}_k = (X_{k1}, \ldots, X_{kP})$ be latent values of the traits at node ν_k on the tree for $k = 1, \ldots, 2N - 1$. For tip nodes $i = 1, \ldots, N$, we posit a stochastic link $p(\mathbf{Y}_i | \mathbf{X}_i)$ where $\mathbf{Y}_i$ is drawn from some distribution parameterized by $\mathbf{X}_i$ and other hyperparameters (see Figure 1). Comparative methods standardly assume that the density $p(\mathbf{Y}_i | \mathbf{X}_i)$ is degenerate at $\mathbf{X}_i$ (i.e. $\mathbf{Y}_i = \mathbf{X}_i$ with probability 1), but we relax this assumption in future sections.

The most common phenotypic model of evolution (Felsenstein 1985) assumes a multivari-
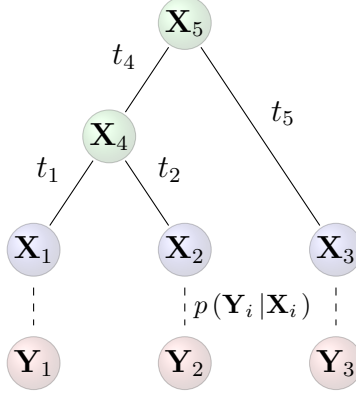
Figure 1: Schematic of diffusion model with stochastic link function. The data $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3)^t$ arise from latent values $\mathbf{X}_i$ at the tips of the tree via the stochastic link function $p(\mathbf{Y}_i | \mathbf{X}_i)$ for $i = 1, \ldots, N$.

ate Brownian diffusion process acts conditionally independently along each branch generating a multivariate normal (MVN) increment,

$$\mathbf{X}_k \sim \text{MVN}\left(\mathbf{X}_{\text{pa}(k)}, t_k \boldsymbol{\Sigma}\right) \text{ for } k = 1, \ldots, 2N - 2, \tag{1}$$

centered around the realized value $\mathbf{X}_{\text{pa}(k)}$ at its parent node and variance proportional to an estimable $P \times P$ positive-definite matrix $\boldsymbol{\Sigma}$. Since the trait values at the root are also unknown, Pybus et al. (2012) suggest further assuming $\mathbf{X}_{2N-1} \sim \text{MVN}\left(\boldsymbol{\mu}_0, \kappa_0^{-1} \boldsymbol{\Sigma}\right)$ with fixed prior mean $\boldsymbol{\mu}_0$ and sample-size κ_0.

2.1 Computation of Observed Data Likelihood

When there are no missing data and under our standard assumption that $p(\mathbf{Y}_i | \mathbf{X}_i)$ is degenerate, integrating out unobserved internal and root node traits leads to a seemingly simple expression for the data likelihood $p(\mathbf{Y} \,|\, \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0)$ (Freckleton 2012; Vrancken et al. 2015). Namely, $\mathbf{Y}$ is matrix-normal (MN) distributed around mean $\mathbf{1}_N \boldsymbol{\mu}_0^t$, with across-row variance $\boldsymbol{\Upsilon} + \kappa_0^{-1} \mathbf{J}_N$ and across-column variance $\boldsymbol{\Sigma}$, where $\mathbf{1}_N$ is a vector of length N populated by ones, $\mathbf{J}_N = \mathbf{1}_N \mathbf{1}_N^t$, and $\boldsymbol{\Upsilon}$ is a deterministic function of $\mathscr{F}$. Specifically, element $\Upsilon_{ii'}$ measures shared evolutionary history and equals the sum of the branch lengths from the root to the most recent common ancestral node of taxa i and i' when $i \neq i'$ or the sum of the branch lengths from the root to taxon i otherwise. For example, in Figure 1, $\Upsilon_{12} = t_4$ and $\Upsilon_{11} = t_1 + t_4$. One can evaluate this highly structured matrix-normal likelihood function with computational complexity $\mathcal{O}(NP^2)$ given the acyclic nature of $\mathscr{F}$. When some data points are missing, however, the observed-data likelihood is no longer matrix-normal and

4

new approaches are needed. This becomes increasingly urgent as the prevalence of missing observations grows with the size of trait data sets. In this context we wish to compute

$$p(\mathbf{Y}^{\text{obs}} \,|\, \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0) = \int p(\mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}} \,|\, \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0) \mathrm{d}\mathbf{Y}^{\text{mis}}, \qquad (2)$$

where $\mathbf{Y}^{\text{obs}}$ and $\mathbf{Y}^{\text{mis}}$ contain the observed and missing trait values, respectively.

The two simplest strategies for calculating the observed-data likelihood are, unfortunately, computationally prohibitive for most large problems. One such solution forfeits the MN structure of the data in favor a simple expression of the observed-data likelihood. This strategy uses the fact that the matrix-normal distribution of $\mathbf{Y}$ can also be expressed as

$$\text{vec}\left[\mathbf{Y} \,|\, \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0\right] \sim \text{MVN}\left(\text{vec}\left[\mathbf{1}_N \boldsymbol{\mu}_0^t\right], \boldsymbol{\Sigma} \otimes \boldsymbol{\Upsilon}\right), \qquad (3)$$

using the Kronecker product $\otimes$. Assuming data are missing at random (Little and Rubin 1987), one can simply remove the rows and columns of $\text{vec}\left[\mathbf{1}_N \boldsymbol{\mu}_0^t\right]$ and $\boldsymbol{\Sigma} \otimes \boldsymbol{\Upsilon}$ corresponding to the missing data and compute the likelihood for this $NP - M'$ dimensional MVN distribution, where M' is the number of missing measurements. This likelihood calculation carries the onerous computational complexity $\mathcal{O}\left((NP - M')^3\right)$. Alternatively, from a Bayesian perspective, one could numerically integrate out the missing data by treating each missing data point as an unknown model parameter and employing MCMC to sample each value. This strategy restores the matrix-normal structure, but requires the likelihood be evaluated each time one samples a missing data point. This results in computation complexity of at least $\mathcal{O}(NP^2M)$, where M is the number of taxa with missing measurements. Because M often scales with N, this method remains prohibitively slow for many data sets with large N. Our goal is to integrate out these missing values analytically using a dynamic programming algorithm in order to bring run time down to a much more manageable $\mathcal{O}(NP^3)$.

2.1.1 Missing Data Definitions and Operations

To develop our algorithm, we first introduce some useful abstractions and notation. At each tip in $\mathscr{F}$, information about each of the P traits comes in one of three forms: a trait value may be directly observed, latent, or completely missing. When directly observed, we posit without loss of generality that the value arises from a normal distribution centered at the observed value with infinite precision. We assume that trait data that arise from latent values are jointly multivariate normally distributed about the unknown latent values with known or estimable precision. Finally, a completely missing value arises also without loss of generality from a normal distribution centered at 0 with zero precision. To formalize this,

for tip $i = 1, \ldots, N$, we construct a permutation matrix $\mathbf{C}_i$ that groups traits in directly observed, latent, and completely missing order and populate a pseudo-precision matrix

$$\mathbf{P}_i = \mathbf{C}_i \operatorname{diag}\left[\infty\mathbf{I}, \mathbf{R}_i, 0\mathbf{I}\right] \mathbf{C}_i^t, \tag{4}$$

where $\operatorname{diag}\left[\cdot\right]$ is a function that arranges its constituent elements into block-diagonal form and $\mathbf{R}_i$ is the latent block precision. Note that any block may be 0-dimensional. This construction arbitrarily forces off-diagonal elements of $\mathbf{P}_i$ involving directly observed and completely missing traits to equal 0 and plays an important role in simplifying computations.

We additionally define a series of operations that we will find useful for defining this algorithm. We define the pseudo-inverse

$$\mathbf{P}_i^- = \mathbf{C}_i \operatorname{diag}\left[0\mathbf{I}, \mathbf{R}_i^{-1}, \infty\mathbf{I}\right] \mathbf{C}_i^t. \tag{5}$$

We define the pseudo-determinant $\hat{\operatorname{det}}\left(\right)$ as the product of the non-zero singular values. We also define the matrix $\boldsymbol{\delta}_i = \operatorname{diag}\left[\delta_{i1}, \ldots, \delta_{iP}\right]$ for $i = 1, \ldots, N$, where δ_{ij} is an indicator variable which takes a value of 1 if trait Y_{ij} is observed or latent and 0 if it is missing. Lastly, we define the possibly degenerate multivariate normal density function

$$\log\hat{\phi}(\mathbf{z}; \boldsymbol{\mu}, \mathbf{P}) = \frac{1}{2}\log\hat{\operatorname{det}}\left(\mathbf{P}\right) - \frac{\operatorname{rank}\left(\mathbf{P}\right)}{2}\log 2\pi - \frac{1}{2}\left(\mathbf{z} - \boldsymbol{\mu}\right)^t \mathbf{P}\left(\mathbf{z} - \boldsymbol{\mu}\right),$$

for some argument $\mathbf{z}$, mean $\boldsymbol{\mu}$ and precision $\mathbf{P}$ of appropriate dimensions.

2.1.2 Post-Order Observed Data Likelihood Algorithm

Our goal is to efficiently compute the likelihood $p(\mathbf{Y}^{\mathrm{obs}} \,|\, \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0)$. Following from Pybus et al. (2012), we perform a post-order traversal where we calculate the observed-data partial likelihood $p(\mathbf{Y}^{\mathrm{obs}}_{\lfloor k \rfloor} \,|\, \mathbf{X}_k, \boldsymbol{\Sigma}, \mathscr{F})$ at each node ν_k where $\mathbf{Y}^{\mathrm{obs}}_{\lfloor k \rfloor}$ is the observed data restricted to all descendants of node k on the tree. For example, in Figure 1, $\mathbf{Y}^{\mathrm{obs}}_{\lfloor 4 \rfloor} = \{\mathbf{Y}^{\mathrm{obs}}_1, \mathbf{Y}^{\mathrm{obs}}_2\}$.

We posit that, given an appropriate stochastic link function $p\left(\mathbf{Y}_i \,|\, \mathbf{X}_i\right)$, we can express the observed-data partial likelihood as

$$p(\mathbf{Y}^{\mathrm{obs}}_{\lfloor k \rfloor} \,|\, \mathbf{X}_k, \boldsymbol{\Sigma}, \mathscr{F}) = r_k \hat{\phi}(\mathbf{X}_k; \mathbf{m}_k, \mathbf{P}_k), \tag{6}$$

for all nodes $k = 1, \ldots, 2N - 1$ and some remainder r_k, mean $\mathbf{m}_k$, and precision $\mathbf{P}_k$. Given a parent node ℓ with children j and k, let us assume we can express the observed-data

likelihood of $\mathbf{Y}_{\lfloor j \rfloor}^{\mathrm{obs}}$ and $\mathbf{Y}_{\lfloor k \rfloor}^{\mathrm{obs}}$ as in Equation 6. Conditioning on $\mathbf{X}_\ell$, we can compute

$$p(\mathbf{Y}_{\lfloor \ell \rfloor}^{\mathrm{obs}} \mid \mathbf{X}_\ell, \boldsymbol{\Sigma}, \mathscr{F}) = p(\mathbf{Y}_{\lfloor j \rfloor}^{\mathrm{obs}} \mid \mathbf{X}_\ell, \boldsymbol{\Sigma}, \mathscr{F}) p(\mathbf{Y}_{\lfloor k \rfloor}^{\mathrm{obs}} \mid \mathbf{X}_\ell, \boldsymbol{\Sigma}, \mathscr{F}) \tag{7}$$

as $\mathbf{Y}_{\lfloor j \rfloor}^{\mathrm{obs}}$ and $\mathbf{Y}_{\lfloor k \rfloor}^{\mathrm{obs}}$ are conditionally independent given $\mathbf{X}_\ell$. Using Equations 1 and 6, we form

$$p(\mathbf{Y}_{\lfloor j \rfloor}^{\mathrm{obs}} \mid \mathbf{X}_\ell, \boldsymbol{\Sigma}, \mathscr{F}) = \int p(\mathbf{Y}_{\lfloor j \rfloor}^{\mathrm{obs}} \mid \mathbf{X}_j, \boldsymbol{\Sigma}, \mathscr{F}) p(\mathbf{X}_j \mid \mathbf{X}_\ell, \boldsymbol{\Sigma}, \mathscr{F}) \mathrm{d}\mathbf{X}_j = r_j \hat{\phi}(\mathbf{X}_\ell; \mathbf{m}_j, \mathbf{P}_j^\star), \tag{8}$$

where the branch-deflated pseudo-precision $\mathbf{P}_j^\star = \left(\mathbf{P}_j^- + t_j \boldsymbol{\delta}_j \boldsymbol{\Sigma} \boldsymbol{\delta}_j\right)^-$. See Supplementary Information (SI) Section 1 for details on computing this pseudo-inverse. We use the results of Equation 8 in Equation 7 to compute the partial log-likelihood

$$\begin{aligned}
\log p(\mathbf{Y}_{\lfloor \ell \rfloor}^{\mathrm{obs}} \mid \mathbf{X}_\ell, \boldsymbol{\Sigma}, \mathscr{F}) &= \log r_j + \log r_k + \log \hat{\phi}(\mathbf{X}_\ell; \mathbf{m}_j, \mathbf{P}_j^\star) + \log \hat{\phi}(\mathbf{X}_\ell; \mathbf{m}_k, \mathbf{P}_k^\star) \\
&= \log r_\ell + \log \hat{\phi}(\mathbf{X}_\ell; \mathbf{m}_\ell, \mathbf{P}_\ell),
\end{aligned} \tag{9}$$

where $\mathbf{P}_\ell = \mathbf{P}_j^\star + \mathbf{P}_k^\star$, $\mathbf{m}_\ell$ is a solution to $\mathbf{P}_\ell \mathbf{m}_\ell = \mathbf{P}_j^\star \mathbf{m}_j + \mathbf{P}_k^\star \mathbf{m}_k$, and

$$\begin{aligned}
\log r_\ell = \log r_j + \log r_k &+ \frac{1}{2} \log \hat{\det}(\mathbf{P}_j^\star) + \frac{1}{2} \log \hat{\det}(\mathbf{P}_k^\star) - \frac{\Delta_{jk\ell}}{2} \log 2\pi \\
&- \frac{1}{2} \log \hat{\det}(\mathbf{P}_\ell) - \frac{1}{2}\left(\mathbf{m}_j^t \mathbf{P}_j^\star \mathbf{m}_j + \mathbf{m}_k^t \mathbf{P}_k^\star \mathbf{m}_k - \mathbf{m}_\ell^t \mathbf{P}_\ell \mathbf{m}_\ell\right).
\end{aligned} \tag{10}$$

Note that the change of informative dimensions $\Delta_{jk\ell} = \mathrm{rank}(\mathbf{P}_j^\star) + \mathrm{rank}(\mathbf{P}_k^\star) - \mathrm{rank}(\mathbf{P}_\ell)$. We update $\boldsymbol{\delta}_\ell = \boldsymbol{\delta}_j \vee \boldsymbol{\delta}_k$, where $\vee$ is the element-wise "logical or" operation.

Our algorithm initializes r_i, $\mathbf{m}_i$, and $\mathbf{P}_i$ such that $p(\mathbf{Y}_i^{\mathrm{obs}} \mid \mathbf{X}_i) = r_i \hat{\phi}(\mathbf{X}_i; \mathbf{m}_i, \mathbf{P}_i)$ at the tips of the tree. For the standard assumption that $\mathbf{Y}_i = \mathbf{X}_i$, we have $r_i = 1$, $\mathbf{m}_i = \mathbf{C}_i [\mathbf{Y}_i^{\mathrm{obs}}, \mathbf{0}]$, and $\mathbf{P}_i = \mathbf{C}_i \mathrm{diag}[\infty \mathbf{I}, 0\mathbf{I}] \mathbf{C}_i^t$. We perform a post-order traversal of the tree computing $\mathbf{m}_\ell, \mathbf{P}_\ell$, and r_ℓ for internal nodes $\ell = N+1, \ldots, 2N-2$ using the already-computed node remainders, means, and precisions for their respective child nodes. At the root, $\mathbf{Y}_{\lfloor 2N-1 \rfloor}^{\mathrm{obs}} = \mathbf{Y}^{\mathrm{obs}}$ and we return the observed-data log-likelihood

$$\begin{aligned}
p(\mathbf{Y}^{\mathrm{obs}} \mid \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0) &= \int p(\mathbf{Y}^{\mathrm{obs}} \mid \mathbf{X}_{2N-1}, \boldsymbol{\Sigma}, \mathscr{F}) p(\mathbf{X}_{2N-1} \mid \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \kappa_0) \mathrm{d}\mathbf{X}_{2N-1} \\
&= \int r_{2N-1} \hat{\phi}(\mathbf{X}_{2N-1}; \mathbf{m}_{2N-1}, \mathbf{P}_{2N-1}) \hat{\phi}(\mathbf{X}_{2N-1}; \boldsymbol{\mu}_0, \kappa_0 \boldsymbol{\Sigma}^-) \mathrm{d}\mathbf{X}_{2N-1} \\
&= r_{\mathrm{full}} \int \hat{\phi}(\mathbf{X}_{2N-1}; \mathbf{m}_{\mathrm{full}}, \mathbf{P}_{\mathrm{full}}) \mathrm{d}\mathbf{X}_{2N-1},
\end{aligned} \tag{11}$$

where $\mathbf{P}_{\text{full}} = \mathbf{P}_{2N-1} + \kappa_0 \boldsymbol{\Sigma}^{-1}$ and $\mathbf{m}_{\text{full}} = \mathbf{P}_{\text{full}}^{-1} \left(\mathbf{P}_{2N-1}\mathbf{m}_{2N-1} + \kappa_0 \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 \right)$. The integral evaluates to one, leaving the observed-data log-likelihood

$$
\begin{aligned}
\log p(\mathbf{Y}^{\text{obs}} \,|\, \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0) &= \log r_{\text{full}} \\
&= \log r_{2N-1} - \frac{\operatorname{rank}\left(\mathbf{P}_{2N-1}\right)}{2} \log 2\pi \\
&\quad + \frac{1}{2} \log \hat{\det}\left(\mathbf{P}_{2N-1}\right) + \frac{1}{2} \log \hat{\det}\left(\kappa_0 \boldsymbol{\Sigma}^{-1}\right) - \frac{1}{2} \log \hat{\det}\left(\mathbf{P}_{\text{full}}\right) \\
&\quad - \frac{1}{2} \left(\mathbf{m}_{2N-1}^t \mathbf{P}_{2N-1} \mathbf{m}_{2N-1} + \kappa_0 \boldsymbol{\mu}_0^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 - \mathbf{m}_{\text{full}}^t \mathbf{P}_{\text{full}} \mathbf{m}_{\text{full}} \right).
\end{aligned}
\tag{12}
$$

This tree traversal visits each node in $\mathscr{F}$ exactly once and inverts a $P \times P$ matrix each time, resulting in an overall computational complexity of $\mathcal{O}(NP^3)$ for each likelihood evaluation.

2.2 Inference

The primary parameter of scientific interest is the diffusion variance $\boldsymbol{\Sigma}$. We are also often interested in additional hyper-parameters $\boldsymbol{\theta}$ related to the stochastic link function $p\left(\mathbf{Y}_i \,|\, \mathbf{X}_i\right)$. In cases where the tree structure is unknown, we use sequence data $\mathbf{S}$ to simultaneously infer $\mathscr{F}$. As such, from a Bayesian perspective, we are interested in approximating

$$
p(\boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\theta} \,|\, \mathbf{Y}^{\text{obs}}, \mathbf{S}) \propto p(\mathbf{Y}^{\text{obs}} \,|\, \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\theta}) \, p(\mathscr{F}, \mathbf{S}) p(\boldsymbol{\Sigma}) p(\boldsymbol{\theta}),
\tag{13}
$$

for inference. We place a $\operatorname{Wishart}_P\left(\boldsymbol{\Lambda}_0, \nu\right)$ prior on $\boldsymbol{\Sigma}^{-1}$, where $\boldsymbol{\Lambda}_0$ is a $P \times P$ rate matrix. The prior on $\boldsymbol{\theta}$ depends the problem of interest, and there are many ways to specify $p(\mathscr{F}, \mathbf{S})$ (see Suchard et al. 2018). To approximate the posterior distributions via MCMC simulation, we apply a random scan Metropolis-within-Gibbs (Liu et al. 1995) approach by which we sample parameter blocks one at a time at random from their full conditional distribution.

Let $\mathbf{X} = (\mathbf{X}_1^t, \cdots, \mathbf{X}_N^t)^t$ be the latent trait values at the tips of the phylogeny. The conjugate $\operatorname{Wishart}_P\left(\boldsymbol{\Lambda}_0, \nu\right)$ prior on $\boldsymbol{\Sigma}^{-1}$ implies that

$$
\begin{aligned}
\boldsymbol{\Sigma}^{-1} \,|\, \mathbf{X}, \mathscr{F}, \boldsymbol{\mu}_0, \kappa_0, \nu, \boldsymbol{\Lambda}_0 \sim \\
\operatorname{Wishart}_P \left[\boldsymbol{\Lambda}_0 + \left(\mathbf{X} - \mathbf{J}_N \boldsymbol{\mu}_0^t\right)^t \left(\boldsymbol{\Upsilon} + \frac{1}{\kappa_0} \mathbf{J}_N \right)^{-1} \left(\mathbf{X} - \mathbf{J}_N \boldsymbol{\mu}_0^t\right), \nu + N \right].
\end{aligned}
\tag{14}
$$

We apply the post-order computation method proposed by Ho and Ané (2014) to compute $\left(\mathbf{X} - \mathbf{J}_N \boldsymbol{\mu}_0^t\right)^t \left(\boldsymbol{\Upsilon} + \frac{1}{\kappa_0} \mathbf{J} \right)^{-1} \left(\mathbf{X} - \mathbf{J}_N \boldsymbol{\mu}_0^t\right)$, which has computational complexity $\mathcal{O}(NP^2)$. When $\mathbf{X}$ are known (i.e. when there are no missing values and $p\left(\mathbf{Y}_i \,|\, \mathbf{X}_i\right)$ is degenerate at

$\mathbf{X}_i$), we can sample from the distribution in Equation 14 immediately without any additional steps. However, if either assumption is violated, we must first draw from the full conditional distribution of $\mathbf{X}$ via the data augmentation algorithm described below.

2.2.1 Pre-Order Missing Data Augmentation Algorithm

To sample jointly from the full conditional of $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^t$, we draw on the calculations made in Section 2.1.2 and perform a pre-order traversal of the tree. Note that we omit explicit dependence on the parameters $\boldsymbol{\Sigma}, \mathscr{F}$, and $\boldsymbol{\theta}$ in all calculations below for clarity. Starting at the root, $\mathbf{X}_{2N-1}$, we draw from $\mathbf{X}_{2N-1} | \mathbf{Y}^{\mathrm{obs}}, \boldsymbol{\mu}_0, \kappa_0$. Using Bayes' rule and Equation 11, we see that

$$
\begin{aligned}
p(\mathbf{X}_{2N-1} \,|\, \mathbf{Y}^{\mathrm{obs}}, \boldsymbol{\mu}_0, \kappa_0) &\propto p(\mathbf{Y}^{\mathrm{obs}} \,|\, \mathbf{X}_{2N-1}) p(\mathbf{X}_{2N-1} \,|\, \boldsymbol{\mu}_0, \kappa_0) \\
&\propto \hat{\phi}(\mathbf{X}_{2N-1}; \mathbf{m}_{\mathrm{full}}, \mathbf{P}_{\mathrm{full}}), \text{ which implies that} \quad (15) \\
\mathbf{X}_{2N-1} | \mathbf{Y}^{\mathrm{obs}}, \boldsymbol{\mu}_0, \kappa_0 &\sim \mathrm{MVN}(\mathbf{m}_{\mathrm{full}}, \mathbf{P}_{\mathrm{full}}).
\end{aligned}
$$

After sampling the root traits from their full conditional, we continue the traversal of the tree where we sample each node $\mathbf{X}_j$ conditional on its (previously sampled) parent node $\mathbf{X}_{\mathrm{pa}(j)}$ and the observed data below node j $\mathbf{Y}^{\mathrm{obs}}_{\lfloor j \rfloor}$ for $j = 1, \ldots, 2N-2$. For the internal nodes, we compute $p(\mathbf{X}_j \,|\, \mathbf{Y}^{\mathrm{obs}}_{\lfloor j \rfloor}, \mathbf{X}_{\mathrm{pa}(j)})$ as follows:

$$
\begin{aligned}
p(\mathbf{X}_j \,|\, \mathbf{Y}^{\mathrm{obs}}_{\lfloor j \rfloor}, \mathbf{X}_{\mathrm{pa}(j)}) &\propto p(\mathbf{Y}^{\mathrm{obs}}_{\lfloor j \rfloor} \,|\, \mathbf{X}_j) p(\mathbf{X}_j \,|\, \mathbf{X}_{\mathrm{pa}(j)}) \\
&\propto \hat{\phi}(\mathbf{X}_j; \mathbf{m}_j, \mathbf{P}_j) \, \hat{\phi}(\mathbf{X}_j; \mathbf{X}_{\mathrm{pa}(j)}, (t_j \boldsymbol{\Sigma})^{-1}) \quad (16) \\
&\propto \hat{\phi}(\mathbf{X}_j; \mathbf{n}_j, \mathbf{Q}_j)
\end{aligned}
$$

where $\mathbf{Q}_j = \mathbf{P}_j + (t_j \boldsymbol{\Sigma})^{-1}$ and $\mathbf{n}_j = \mathbf{Q}_j^{-1}(\mathbf{P}_j \mathbf{m}_j + (t_j \boldsymbol{\Sigma})^{-1} \mathbf{X}_{\mathrm{pa}(j)})$. This implies $\mathbf{X}_j | \mathbf{Y}^{\mathrm{obs}}_{\lfloor j \rfloor}, \mathbf{X}_{\mathrm{pa}(j)} \sim \mathrm{MVN}(\mathbf{n}_j, \mathbf{Q}_j)$, and we sample $\mathbf{X}_j$ from this distribution.

At the tips, we employ one of two techniques depending on the specific model. Under our standard assumption (i.e. $\mathbf{X}_i = \mathbf{Y}_i$ with probability 1), we partition the precision $\boldsymbol{\Sigma}^{-1}$ and trait values $\mathbf{X}_i$ and $\mathbf{X}_{\mathrm{pa}(i)}$ such that

$$
\boldsymbol{\Sigma}^{-1} = \mathbf{C}_i \begin{pmatrix} \mathbf{S}_i^{\mathrm{obs}} & \mathbf{S}_i^{\mathrm{om}} \\ \mathbf{S}_i^{\mathrm{mo}} & \mathbf{S}_i^{\mathrm{mis}} \end{pmatrix} \mathbf{C}_i^t, \quad \mathbf{X}_i = \mathbf{C}_i \begin{pmatrix} \mathbf{X}_i^{\mathrm{obs}} \\ \mathbf{X}_i^{\mathrm{mis}} \end{pmatrix}, \quad \text{and} \quad \mathbf{X}_{\mathrm{pa}(i)} = \mathbf{C}_i \begin{pmatrix} \mathbf{X}_{\mathrm{pa}(i)}^{\mathrm{obs}} \\ \mathbf{X}_{\mathrm{pa}(i)}^{\mathrm{mis}} \end{pmatrix} \quad (17)
$$

and draw from $\mathbf{X}_i^{\mathrm{mis}} | \mathbf{Y}_i^{\mathrm{obs}}, \mathbf{X}_{\mathrm{pa}(i)} \sim \mathrm{MVN}\left(\mathbf{X}_{\mathrm{pa}(i)}^{\mathrm{mis}} + \mathbf{S}_i^{\mathrm{mis}-1} \mathbf{S}_i^{\mathrm{mo}}\left(\mathbf{X}_{\mathrm{pa}(i)}^{\mathrm{obs}} - \mathbf{X}_i^{\mathrm{obs}}\right), \frac{1}{t_i} \mathbf{S}_i^{\mathrm{mis}}\right)$ for $i = 1, \ldots, N$. For cases where $p(\mathbf{Y}_i | \mathbf{X}_i)$ is non-degenerate, we simply use Equation 16 to sample from $\mathbf{X}_i | \mathbf{Y}_i^{\mathrm{obs}}, \mathbf{X}_{\mathrm{pa}(i)}$. Once we have sampled $\mathbf{X} | \mathbf{Y}^{\mathrm{obs}}, \boldsymbol{\Sigma}, \mathscr{F}, \boldsymbol{\theta}$, we can draw from the full

conditional distribution of $\mathbf{\Sigma}^{-1}$ via Equation 14. This pre-order data augmentation procedure requires a single $P \times P$ matrix inversion at each of the $2N - 1$ nodes in the tree, resulting in overall computational complexity $\mathcal{O}(NP^3)$.

3. MODEL EXTENSION: RESIDUAL VARIANCE

We extend the MBD model of phenotypic evolution to include multivariate normal residual variance at each of the tips. Under this model, we assume

$$p(\mathbf{Y}_i \,|\mathbf{X}_i) = \hat{\phi}(\mathbf{Y}_i; \mathbf{X}_i, \mathbf{\Gamma}) \ \text{for } i = 1, \ldots, N, \tag{18}$$

where $\mathbf{\Gamma}$ is a $P \times P$ precision matrix. Under this model, the vectorization of $\mathbf{Y}$ is MVN-distributed with $NP \times NP$ variance-covariance matrix $\mathbf{\Sigma} \otimes (\mathbf{\Upsilon} + \kappa_0^{-1}\mathbf{J}_N) + \mathbf{\Gamma}^{-1} \otimes \mathbf{I}_N$ where $\mathbf{I}_N$ is the $N \times N$ identity matrix. Unlike the case where $\mathbf{Y}_i = \mathbf{X}_i$, $\mathbf{Y}$ cannot be expressed as matrix-normal even in the data-complete case because the variance cannot be expressed as the Kronecker product of two matrices. As such, our post-order likelihood computation algorithm is useful for this extended model, even when there are no missing data points.

3.1 Inference of Residual Variance

Similar to our inference of $\mathbf{\Sigma}$ in the diffusion process, we place a conjugate Wishart $(\mathbf{\Lambda}_s, \nu_s)$ prior on $\mathbf{\Gamma}$ using the rate parameterization. This yields the full conditional distribution

$$\mathbf{\Gamma} \,|\mathbf{Y}, \mathbf{X} \sim \text{Wishart}_P \left(\mathbf{\Lambda}_s + (\mathbf{Y} - \mathbf{X})^t (\mathbf{Y} - \mathbf{X}), \nu_s + N\right). \tag{19}$$

Because $\mathbf{X}$ is latent in this model, each time we update $\mathbf{\Gamma}$ we first draw from the full conditional posterior of $\mathbf{X}$ using the algorithm described in Section 2.2.1. For cases where $\mathbf{Y}$ is not completely observed, we must perform an additional data augmentation step where we draw from $\mathbf{Y}^{\text{mis}}\,|\mathbf{Y}^{\text{obs}}, \mathbf{X}, \mathbf{\Gamma}$. To do this, we decompose the sampling precision matrix into blocks such that

$$\mathbf{\Gamma} = \mathbf{C}_i \begin{pmatrix} \mathbf{\Gamma}_i^{\text{obs}} & \mathbf{\Gamma}_i^{\text{mo}t} \\ \mathbf{\Gamma}_i^{\text{mo}} & \mathbf{\Gamma}_i^{\text{mis}} \end{pmatrix} \mathbf{C}_i^t \ \text{for } i = 1, \ldots, N. \tag{20}$$

From Equation 18, we see that

$$p(\mathbf{Y}_i^{\text{mis}} \,|\mathbf{Y}_i^{\text{obs}}, \mathbf{X}_i, \mathbf{\Gamma}) = \hat{\phi}\left(\mathbf{Y}_i^{\text{mis}}; \mathbf{X}_i^{\text{mis}} + \mathbf{\Gamma}_i^{\text{mis}\,-1}\mathbf{\Gamma}_i^{\text{mo}}(\mathbf{X}_i^{\text{obs}} - \mathbf{Y}_i^{\text{obs}}), \mathbf{\Gamma}_i^{\text{mis}}\right). \tag{21}$$

As such, we can directly sample $\mathbf{Y}_i^{\text{mis}}$ from its full conditional above and update $\mathbf{Y}_i = \mathbf{C}_i [\mathbf{Y}_i^{\text{obs}}, \mathbf{Y}_i^{\text{mis}}]^t$ for $i = 1, \ldots, N$. This process also has computational complexity $\mathcal{O}(NP^3)$.

Note that we can draw from the joint full conditional of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ by performing a single pre-order data augmentation where we draw from $p(\mathbf{X}, \mathbf{Y}^{\mathrm{mis}} \,|\, \boldsymbol{\Sigma}, \boldsymbol{\Gamma})$ and subsequently draw from $p(\boldsymbol{\Sigma}, \boldsymbol{\Gamma} \,|\, \mathbf{X}, \mathbf{Y}^{\mathrm{mis}}) = p(\boldsymbol{\Sigma} \,|\, \mathbf{X}) p(\boldsymbol{\Gamma} \,|\, \mathbf{X}, \mathbf{Y})$. These distributions are conditionally independent due to the fact that $\mathbf{X}$ and $\mathbf{X} - \mathbf{Y}$ are independent by construction. This procedure effectively halves the computation time as we only need to perform a single post-order likelihood computation/pre-order data augmentation step to sample both $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$, rather than each time we sample one.

3.2 Heritability Statistic

The residual variance extension enables us to estimate phenotypic heritability over evolutionary time. We use a definition analogous to the broad-sense heritability in statistical genetics (see Visscher et al. 2008). Namely, we seek to quantify the proportion of variance in a trait attributable to the Brownian diffusion process on the phylogeny (as opposed to the residual variance). Note that we are primarily interested in heritability in the HIV example below, for which we use data from a recent paper by Blanquart et al. (2017). As such, we use a multivariate generalization of the heritability statistic from that paper. Specifically, we estimate phylogenetic heritability by taking the expectation of the empirical sample variance under our extended model. We define the $P \times P$ empirical covariance matrix as

$$\mathbf{S}^2(\mathbf{Y}) = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{Y}_i - \bar{\mathbf{y}}) (\mathbf{Y}_i - \bar{\mathbf{y}})^t = \frac{1}{N} \left(\mathbf{Y} - \bar{\mathbf{Y}} \right)^t \left(\mathbf{Y} - \bar{\mathbf{Y}} \right), \qquad (22)$$

where $\bar{\mathbf{y}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{Y}_i = \frac{1}{N} \mathbf{Y}^t \mathbf{1}_N$ and $\bar{\mathbf{Y}} = \mathbf{1}_N \bar{\mathbf{y}}^t = \frac{1}{N} \mathbf{J}_N \mathbf{Y}$. The expectation of this quantity reduces to the following expression (see SI Section 2 for details):

$$\mathbb{E}\left[\mathbf{S}^2(\mathbf{Y}) \right] = \frac{N-1}{N} \boldsymbol{\Gamma}^{-1} + \left(\frac{1}{N} \mathrm{tr}\left[\boldsymbol{\Upsilon} \right] - \frac{1}{N^2} \mathbf{1}_N^t \boldsymbol{\Upsilon} \mathbf{1}_N \right) \boldsymbol{\Sigma}. \qquad (23)$$

Because $\mathbb{E}\left[\mathbf{S}^2(\mathbf{Y}) \right]$ is a linear combination of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}^{-1}$, we propose the $P \times P$ heritability matrix $\mathbf{H} = \{ h_{kl} \}$ with entries

$$h_{kl} = \frac{c_\sigma \Sigma_{kl}}{\sqrt{\left(c_\sigma \Sigma_{kk} + c_\gamma \Gamma_{kk}^{-1} \right) \left(c_\sigma \Sigma_{ll} + c_\gamma \Gamma_{ll}^{-1} \right)}}, \qquad (24)$$

where $c_\sigma = \frac{1}{N} \mathrm{tr}\left[\boldsymbol{\Upsilon} \right] - \frac{1}{N^2} \mathbf{1}_N^t \boldsymbol{\Upsilon} \mathbf{1}_N$ and $c_\gamma = \frac{N-1}{N}$. The diagonal entries $h_{kk} = h_k^2$ represent the marginal phylogenetic heritability of each trait, and the off-diagonal entries represent pairwise co-heritability (Falconer 1960, chap. 19) between traits. Note that naive computation

11

of c_σ relies on constructing $\boldsymbol{\Upsilon}$ that has complexity $\mathcal{O}(N^2)$. SI Section 3 describes a novel post-order algorithm that computes c_σ in $\mathcal{O}(N)$ time without explicitly constructing $\boldsymbol{\Upsilon}$.

While the breadth of research in heritability is extensive across both statistical genetics and phylogenetics (see in particular the recent paper by Mitov and Stadler 2018), we choose the same heritability statistic as used by Blanquart et al. (2017) for direct comparison with their analysis. That being said, our methods could be readily adapted to approximate the posterior distribution of several of the alternative heritability statistics presented in Mitov and Stadler (2018). Additionally, our pre-order data augmentation procedure allows us to generate samples directly from the posterior of the latent trip traits $\mathbf{X}$, from which we can directly compute the genetic covariance $\mathbf{S}^2(\mathbf{X})$ rather than relying on expectations.

4. RESEARCH MATERIALS

We have implemented these methods in the development version of BEAST (Drummond et al. 2012). The XML files for all analyses and instructions for running them are available at `https://github.com/suchard-group/missing_traits_paper`.

5. COMPUTATIONAL EFFICIENCY

Our method dramatically increases computational efficiency over the current best-practice method. This latter procedure, developed by Cybis et al. (2015), treats the missing and latent values of $\mathbf{X}$ as unknown parameters and numerically integrates them out by placing a Gibbs sampler on each tip $\mathbf{X}_i$ that draws from its full conditional distribution $p\left(\mathbf{X}_i \,\middle|\, \mathbf{Y}_i, \mathbf{X}_{\lceil i\rceil}\right)$ for $i = 1, \ldots, N$ where $\mathbf{X}_{\lceil i\rceil} = \mathbf{X} \backslash \mathbf{X}_i$. Because the full conditional distribution of $\mathbf{X}_i$ relies on the other missing and latent values in $\mathbf{X}$, we sample each tip individually. The advantage of this is that the likelihood calculation, the Gibbs sampler of the diffusion variance $\boldsymbol{\Sigma}$, and the data augmentation procedure for each tip all have complexity $\mathcal{O}(NP^2)$ rather than our $\mathcal{O}(NP^3)$. As such, this numerical integration procedure has overall complexity $\mathcal{O}(MNP^2)$ where M is the number of tips with missing or latent values. For any extended model where $p\left(\mathbf{Y}_i \,\middle|\, \mathbf{X}_i\right)$ is not degenerate at $\mathbf{X}_i$, all values of $\mathbf{X}$ are latent and $M = N$.

We formalize our comparison by computing the median and minimum effective sample size (ESS) per hour for all parameters of interest under both our analytical integration method and the sampling method discussed above. We also compute the ESS per sample and samples per hour to understand how our improved method influences both the autocorrelation between MCMC samples and the amount of computation work required to generate a single draw from the posterior. We define the number of samples as the number of states

Table 1: Algorithmic improvement. We report MCMC sampling efficiency through effective sample size (ESS) that shows both a decrease in autocorrelation (as shows by ESS / Sample) and in the overall work required per sample (as shown by Samples / Hour).

Data set	Model	Integration method	ESS/hour		ESS/sample		Samples /hour
			minimum	median	minimum	median	
Mammals	Diffusion only	Analytic	1,200	3,600	0.044	0.13	27,000
		Sampling	3	9.8	0.0043	0.014	700
		Speed-up	**410×**	**360×**	**10×**	**9.3×**	**39×**
	Diffusion with residual	Analytic	140	320	0.0062	0.015	22,000
		Sampling	0.38	3	0.00024	0.0018	1,600
		Speed-up	**350×**	**110×**	**26×**	**7.9×**	**14×**
HIV	Diffusion only	Analytic	100,000	220,000	0.31	0.66	320,000
		Sampling	1,500	8,500	0.01	0.057	150,000
		Speed-up	**65×**	**25×**	**30×**	**12×**	**2.2×**
	Diffusion with residual	Analytic	1,600	2,500	0.0061	0.0096	260,000
		Sampling	5.1	8.7	9.9e-5	0.00017	51,000
		Speed-up	**320×**	**290×**	**61×**	**56×**	**5.2×**

in which the MCMC simulation updates the parameters of interest (as opposed to missing trait values). Table 1 presents the results of our efficiency comparisons.

6. APPLICATIONS

6.1 Mammalian Life History

A major task for life history theory is to understand the ecological and evolutionary significance of correlation between life history traits such as age at sexual maturity, the number of offspring per reproductive event, and reproductive lifespan (Roff 2002). Establishing patterns of such correlation grants insight into whether life history variation between individuals, populations or species is consistent with pace-of-life theory (Reynolds 2003; Réale et al. 2010). This theory predicts that 'fast' traits such as early maturity, large broods, small offspring, frequent reproduction and a short lifespan are positively associated with each other as a consequence of organisms pursuing strategies that prioritize either current or future reproduction. Existing approaches using comparative life history data to investigate fast-slow trait covariation patterns (e.g. mammals: Bielby et al. 2007; hymenoptera: Blackburn 1991; lizards: Clobert et al. 1998; birds: Sæther and Bakke 2000; plants: Salguro-Gómez 2017;

fish: Wiedmann et al. 2014) generally support the fast-slow hypothesis; however, results are rarely consistent across taxa. This may reflect important taxonomic differences in life history evolution, but there is concern that differences are an artifact of different methodologies (Jeschke and Kokko 2009).

One key limitation is that previous methods have required complete data for each species. As complete measurements across a rich suite of varied life history traits are not yet available for most species, this means that researchers must choose to either reduce the number of traits or reduce the number of species included in analyses. By integrating out missing traits, we resolve this issue and analyze the life history dataset used in Capellini et al. (2015), which is based largely on the final PanTHERIA dataset (Jones et al. 2009), supplemented with measurements from Ernest (2003) and additional sources. Our analysis includes all the variables analyzed by Bielby et al. (2007) (gestation length, weaning age, neonatal body mass, litter size, litter frequency, and age at first birth) plus reproductive lifespan (maximum lifespan minus age at first birth). We include female body mass as a trait rather than analyze size-corrected residuals and log-transform and standardize all traits prior to analysis. The analysis assumes the phylogeny of Fritz et al. (2009) that remains the most complete phylogeny for mammals. In total, 3690 species in the phylogeny have measurement of at least one trait and are included. Table 2 reports the number of species with measurements for each trait. Only 518 species have complete data on all 8 traits; thus the ability to include species with partially missing traits enables inclusion of 208% more measurements.

To estimate the correlation between these traits throughout mammalian evolution, we jointly model them with an MBD process on the tree with residual variance. In this analysis, we are primarily interested in the correlation between traits during the MBD process on the tree and estimate trait correlations from the marginal posterior of Σ. Figure 2 summarizes these findings. Our results are clearly consistent with the fast-slow trait covariation patterns that pace-of-life theory predicts. The 'slow' life history traits (longer gestation, later weaning, larger neonatal body mass, later age at first birth, and longer reproductive lifespan) are all positively correlated with each other and negatively correlated with the two 'fast' life history traits (greater litter size and more frequent litters). All correlations are significant (determined by $< 5\%$ posterior tail probability) with the notable exception of that between litter size and litter frequency. This apparent lack of correlation may be due to the opposing effects of their joint positive correlation with body mass combined with a trade-off between these two traits that life history theory predicts (Stearns 1989). Nevertheless, our results demonstrate that larger animals tend to have slower life history traits, confirming known patterns and reflecting the central role of body size in life history evolution.

We compare the computational efficiency of our method against that of the sampling

Table 2: Missing data summary for all three examples.

Data set	Trait	Number observed	Percent missing
Mammals $N = 3690$	Body mass	3508	4.9%
	Litter size	2538	31.2%
	Gestation length	1427	61.3%
	Weaning age	1253	66.0%
	Litter frequency	1231	66.6%
	Neonatal body mass	1108	70.0%
	Age at first birth	945	74.4%
	Reproductive lifespan	748	79.7%
	Total	**12758**	**56.8%**
Prokaryotes $N = 705$	Cell diameter	690	2.1%
	Cell length	657	6.8 %
	Genome length	563	20.1%
	GC content	563	20.1%
	Coding sequence length	558	20.9%
	Optimal temperature	548	22.2%
	Optimal pH	487	30.9%
	Total	**4066**	**17.6%**
HIV $N = 1536$	GSVL	1536	0.0%
	SPVL	1536	0.0%
	CD4 slope	1102	28.3%
	Total	**4174**	**9.4%**

method using the MBD model both with and without residual variance. Table 1 shows an increase in overall computational efficiency of two orders-of-magnitude as indicated by the change in ESS per hour. Additionally, we see that our method succeeds at reducing both the amount of computational work per MCMC sample (as indicated by the increase in samples per hour) and autocorrelation (as indicated by the increase in ESS per sample).

6.2 Prokaryote evolution

Comparative genomics has greatly assisted in the formulation of prokaryote evolutionary theories. Several such theories have been inspired by and tested through measuring correlation among different phenotypic and genomic traits. For example, the thermal adaptation hypothesis posits that higher GC content is involved in adaptation to high temperatures because it may offer thermostability to genetic material (Bernardi and Bernardi 1986). The
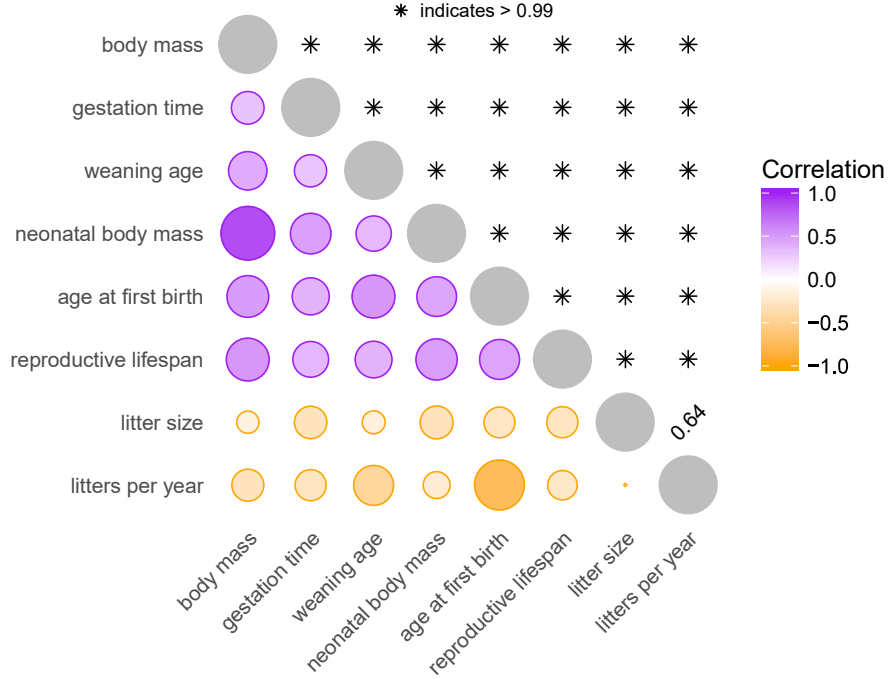
Figure 2: Correlation among mammalian life-history traits. The circles below the diagonal summarize the posterior mean correlation between each pair of traits. Purple represents a positive correlation while orange represents a negative correlation. Circle size and color intensity both represent the absolute value of the correlation. The numbers above the diagonal report the posterior probability that the correlation is of the same sign as its mean.

genome streamlining hypothesis attempts to explain the compactness of prokaryotic genomes through natural selection favoring small genomes (Doolittle and Sapienza 1980; Orgel and Crick 1980; Giovannoni et al. 2014). Sabath et al. (2013) argue that lower cell volume is an adaptive response to high temperature. The field is well-aware of the need to account for phylogenetic relationships when measuring correlation, but statistical analyses generally rely on fixed, poorly resolved trees and simple models of trait evolution.

Here, we estimate correlation among a set of genotypic and phenotypic traits while simultaneously accounting for phylogenetic uncertainty and accommodating complexity in the trait evolutionary process. We construct our data set from a study by Goberna and Verdú (2016), who collated cell diameter, cell length, optimum temperature and pH measurements for a large set of prokaryotes. Prior experience in resolving large, unknown trees suggests that we limit our analysis to less than ∼750 taxa. As such, we include all taxa with three or more measurements and a selection of the taxa with only two measurements in our analysis. For our selection of 705 taxa, we obtain data on genome length, the number of coding sequences, and GC content from the prokaryotes table in NCBI Genome. Table 2 presents the

number of measurements for each trait. We log-transform and standardize all traits (except for GC content which we logit-transform and standardize). To infer the phylogeny, we obtain matching 16S sequences via the ARB software package (Ludwig et al. 2004) that we then align using the SINA Alignment Service (Pruesse et al. 2012) and manually edit.
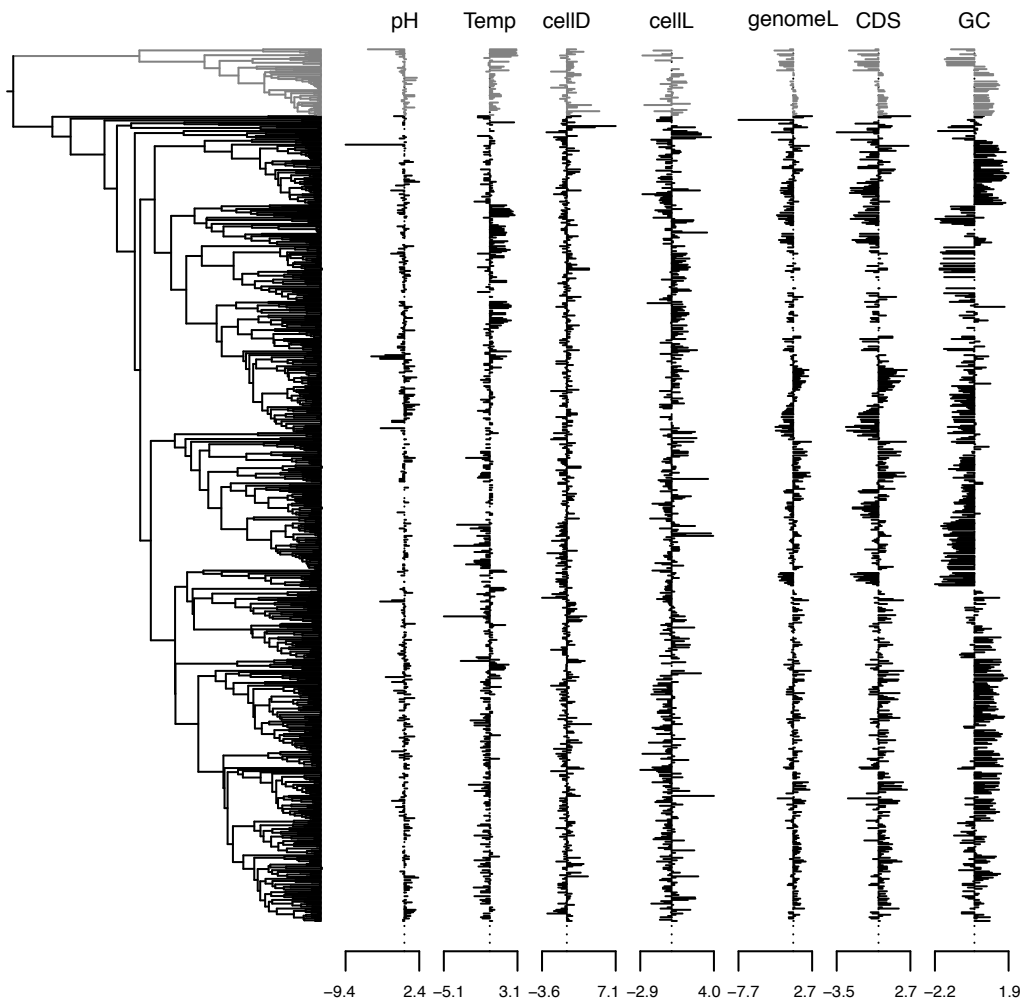

Figure 3: Prokaryote phylogeny and traits. The phylogeny depicts the inferred maximum clade credibility tree. The archaea clade ($N = 54$) and the associated trait measurements are depicted in grey.

Through MCMC simulation, we simultaneously infer the sequence and trait evolutionary process. We model the sequence evolutionary process using a general time-reversible model (Tavaré 1986) with gamma-distributed rate variation among sites (Yang 1994). We use an uncorrelated lognormal relaxed clock to model rate variation among branches (Drummond et al. 2006) and specify a Yule birth prior process on the unknown tree (Gernhard 2008). For the trait evolutionary process, we assume an MBD model with residual variance.

17
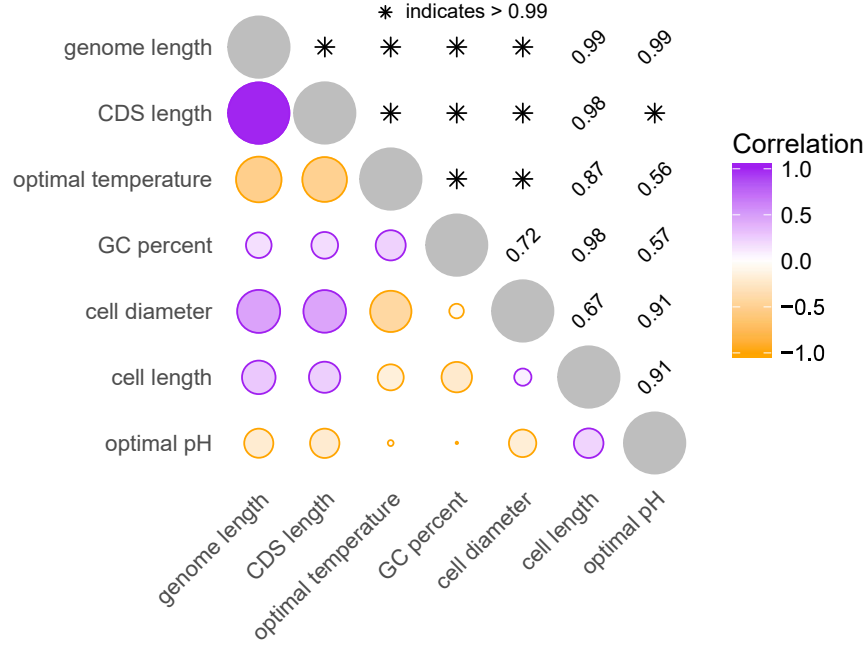
Figure 4: Correlation among prokaryotic growth properties. See Figure 2 caption.

Figure 3 displays our estimated maximum clade credibility phylogeny with associated trait measurements, and Figure 4 presents the phylogenetic correlation between those traits. One notable result is the positive correlation between optimal temperature and GC content (0.22 posterior mean, $[0.08, 0.37]$ 95% highest posterior density interval) that the thermal adaptation hypothesis predicts (Bernardi and Bernardi 1986). Researchers have discussed this hypothesis for years with mixed support (Hurst and Merchant 2001; Musto et al. 2004; Wang et al. 2006; Wu et al. 2012; Sabath et al. 2013; Aptekmann and Nadra 2018). Our analysis, however, includes 435 taxa with measurements for both GC content and optimal growth temperature, making it the largest study we are aware of that accounts for phylogenetic relationships. Interestingly, while cell diameter and cell length are not significantly correlated, they are both positively correlated with genome length. Smaller cells have been associated with smaller genomes in both prokaryotes and unicellular eukaryotes (Shuter et al. 1983; Lynch 2007), but the reasons for this are not fully understood (Dill et al. 2011). We also estimate a relatively strong negative correlation between genome length and optimal temperature (-0.52 $[-0.67, -0.37]$), supporting the genomic streamlining hypothesis during thermal adaptation. Note that we do not compare computation times here, as simultaneous inference of the phylogenetic tree makes the sampling method prohibitively slow.

6.3 HIV-1 virulence

Recent years have witnessed a strong interest in using phylogenetic comparative methods to study the heritability of HIV-1 virulence. Initially, Alizon et al. (2010) employed Pagel's λ (Pagel 1999) to measure the extent to which HIV-1 set-point viral load reflects viral shared evolutionary history in the Swiss HIV Cohort Study (Swiss HIV Cohort Study et al. 2009) patients. A relatively high heritability estimate of set-point viral load, a predictive measure of clinical outcome, motivated others to examine to what extent the viral genotype can control for this trait (e.g. Hodcroft et al. 2014; Vrancken et al. 2015). These efforts have resulted in widely varying estimates, from 6% to 59%, prompting a discussion on the methods used to estimate the heritability of virulence (see Mitov and Stadler 2018; Bertels et al. 2018). Here, we revisit the most comprehensive data set recently analyzed (Blanquart et al. 2017) to determine the extent to which variability in HIV-1 virulence is attributable to viral genetic variation. We focus on the dataset of subtype B viruses from Blanquart et al. (2017) that encompasses 1581 taxa with associated measures of set-point viral load and CD4 cell count decline. We rely on the maximum likelihood phylogeny inferred for this data set, but convert it to a time-measured tree with dated tips using a heuristic procedure (To et al. 2016). A prior examination of the correlation between sampling time and root-to-tip divergence using TempEst (Rambaut et al. 2016) indicated the presence of outliers, most of which can be attributed to a basal lineage in the phylogeny. As the subtyping of the taxa in this basal lineage also was ambiguous (Blanquart, personal communication), we remove this lineage (36 taxa) together with 9 other outlier taxa. We note that this resulted in a time to the most recent common ancestor (TMRCA) estimate of about 1960 that is much more in line with a recent subtype B TMRCA estimate (1967, 95% Bayesian credibility interval of 1963–1970; Worobey et al. 2016) than the estimate including the basal lineage (∼1930).

Two measures of set-point viral load are available for all remaining taxa: (i) one based on a standardized choice of assay on a single sample taken between 6 and 24 months after infection and before the initiation of antiretroviral therapy ("gold standard viral load", GSVL) and (ii) a more classical measure of set-point viral load (SPVL) based on the mean of all log viral loads measured between 6 and 24 months after infection. Figure 5 presents the phylogeny and associated trait values. To estimate heritability of both set-point viral load measures and CD4 slope, we model all three measurements as a multivariate trait in our MBD model with residual variance and approximate the posterior distribution of the heritability statistic **H** via MCMC. Our estimated heritabilities are 0.21 [0.11, 0.3] for GSVL, 0.18 [0.1, 0.26] for SPVL, and 0.16 [0.07, 0.25] for CD4 cell decline. These estimates are consistent with similar estimates reported by Blanquart et al. (2017)

We further asses model fit by assessing predictive performance of GSVL on SPVL. We
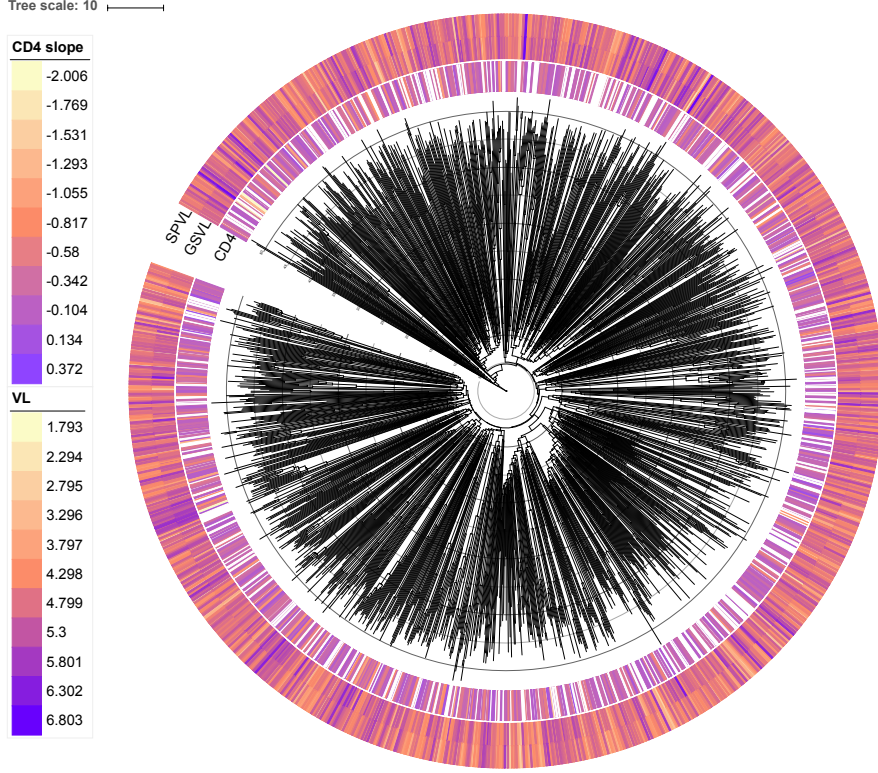
19

Figure 5: HIV-1 phylogeny with associated CD4 slope, SPVL, and GSVL values for each viral host.

omit CD4 slope from our analysis as it is measured concurrently with SPVL. We randomly remove 5% of the SPVL measurements from the data set and consider four different models. We consider both a bivariate case where we assume a multivariate process and a univariate case where we analyze SPVL alone. For both the bivariate and univariate cases, we use the MBD model both with and without the residual variance extension. For each removed SPVL measurement, we compute the mean squared error (MSE) between the predicted and true values. We repeat each analysis 50 times and report the cumulative results in Figure 6, from which two results emerge. First, the MSE of prediction in the bivariate cases are lower than those in the univariate cases. This is unsurprising given the strong correlation between SPVL and GSVL. Second, addition of residual variance to the model results in modestly better prediction of SPVL in both the bivariate and univariate cases. This emphasizes the importance of including model extensions like residual variance in these analyses.

We again demonstrate improvements in computational efficiency (see Table 1). While less dramatic than the mammals example, we still see an order-of-magnitude increase in effective
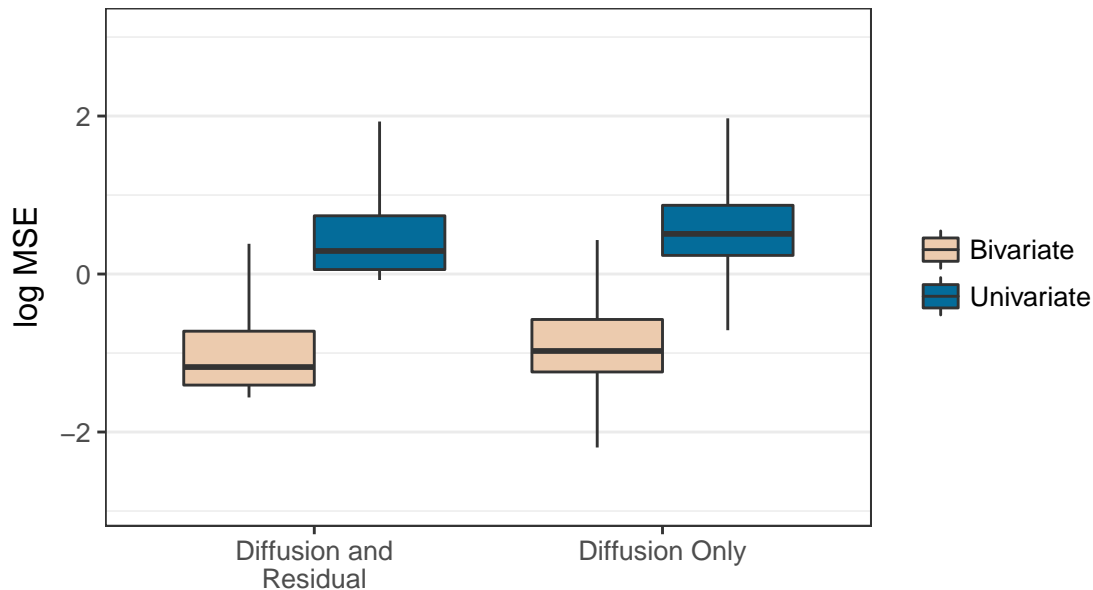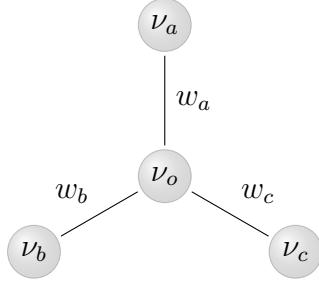
Figure 6: Model predictive performance of HIV set-point viral load. Each box-and-whisker plot depicts the posterior mean-squared-error of prediction under a different model. The boxes represent the interquartile range, while the lines extend to include the 2.5^{th} through 97.5^{th} percentiles. Outliers are omitted.

sample size per hour in the MBD model without residual variance. This attenuation is to be expected, as there are far fewer missing measurements in the HIV data set than the mammal data set. Nevertheless, our method still outperforms the sampling method in the simple MBD model even when only 9.4% of data points are missing. For the model with residual variance, our method outperforms the sampling method by two orders-of-magnitude.

7. DISCUSSION

Oftentimes comparative biologists are interested in phylogenetically adjusted methods for assessing relationships between traits of organisms. However, frequently when the number of taxa grows large the level of missing data increases, making inference challenging. Here, we have developed a method for evaluating the likelihood of observed traits given a tree while integrating out missing values analytically that dramatically outperforms current best-practice methods. In the mammalian data set, with $N = 3690$ and 56.8% missing data, we achieve a minimum effective sample size per hour 410× greater than previous methods. This increase in speed brings computation times down from more than a week to less than an hour. Even in the more tractable HIV data set, with $N = 1536$ and 9.4% missing data, we increase the minimum ESS per hour by a factor of 65. Both increases in speed are due

21

Figure 7: An acyclic graph with nodes $\{\nu_o, \nu_a, \nu_b, \nu_c\}$ and edge weights $\{w_a, w_b, w_c\}$. The covariance matrix $\mathbf{\Lambda} = \{\Lambda_{ij}\}$ is additive on an acyclic graph if each Λ_{ij} is equal to the sum of the shared non-negative edge-weights in the paths from ν_i and ν_j to some origin node. For example, the matrix $\mathbf{M}_1$ is additive for nodes $(\nu_a, \nu_b, \nu_c)^t$ with ν_o at the origin, while the matrix $\mathbf{M}_2$ is additive for nodes $(\nu_o, \nu_b, \nu_c)^t$ with ν_a at the origin.

to an overall decrease in both autocorrelation between MCMC samples and the amount of computational work required per sample. Importantly, this increase in computational efficiency allows for previously intractable analyses on large trees. Specifically, we incorporate residual variance into the model and (in the prokaryotes example) simultaneously infer $\mathbf{\Sigma}$, $\mathbf{\Gamma}$, and an unknown phylogeny $\mathscr{F}$. Further, the residual variance extension is only one of several potential extensions. Other possible extensions could incorporate data sets with repeated measurements at the tips of the tree and factor analyses (Tolkoff et al. 2017).

An important limitation of our and previous methods is that they assume an ignorable missing data mechanism (i.e. that the data are missing at random and that the prior on any model parameters is independent of the missing data mechanism) (Little and Rubin 1987). While these conditions may hold in some comparative biology examples, possible violations abound (e.g. any data set where values below some detection limit are omitted). One potential solution explicitly includes these thresholds in the model and renormalizes the observed-data likelihood appropriately.

Finally, and perhaps most importantly, we propose our method as a special case solution to the long-standing statistical problem involving multivariate normal distributions with missing data. Specifically, our method applies to any MVN distribution with a three-point structured covariance matrix (see Ho and Ané 2014). Intuitively, this condition arises in covariance matrices generated from processes that are additive on an acyclic graph (see Figure 7). This restriction, however, is not overly limiting and applies to a broad range of normal models including multilevel hierarchical models and matrix-normal distributions such as the one we use here. Additionally, our pre-order data augmentation procedure enables $\mathcal{O}(N)$ imputation in these highly structured models. While Allen and Tibshirani (2010) and Glanz and Carvalho (2018) have utilized the EM algorithm (Dempster et al. 1977) to

efficiently perform maximum likelihood imputation in similar problems, our method could serve as an alternative for approaches that base inference on the observed-data likelihood.

SUPPLEMENTARY MATERIAL

Detailed Calculations: Includes SI Sections 1 through 3 (pdf file)

Bibliography

Adams, D. C. (2014). A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. *Evolution 68*(9), 2675–2688.

Alizon, S., V. von Wyl, T. Stadler, R. D. Kouyos, S. Yerly, B. Hirschel, J. Böni, C. Shah, T. Klimkait, H. Furrer, A. Rauch, P. L. Vernazza, E. Bernasconi, M. Battegay, P. Bürgisser, A. Telenti, H. F. Günthard, S. Bonhoeffer, and Swiss HIV Cohort Study (2010, Sep). Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathogens 6*(9), e1001123.

Allen, G. and R. Tibshirani (2010). Transposable regularized covariance models with an application to missing data imputation. *Annals of Applied Statistics 4*, 764–790.

Aptekmann, A. A. and A. D. Nadra (2018). Core promoter information content correlates with optimal growth temperature. *Scientific Reports 8*(1), 1313.

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

Bernardi, G. and G. Bernardi (1986). Compositional constraints and genome evolution. *Journal of Molecular Evolution 24*(1-2), 1–11.

Bertels, F., A. Marzel, G. Leventhal, V. Mitov, J. Fellay, H. F. Günthard, J. Böni, S. Yerly, T. Klimkait, V. Aubert, M. Battegay, A. Rauch, M. Cavassini, A. Calmy, E. Bernasconi, P. Schmid, A. U. Scherrer, V. Müller, S. Bonhoeffer, R. Kouyos, R. R. Regoes, and Swiss HIV Cohort Study (2018, Jan). Dissecting HIV virulence: Heritability of setpoint viral load, CD4+ T-cell decline, and per-parasite pathogenicity. *Molecular Biology and Evolution 35*(1), 27–37.

Bielby, J., G. Mace, O. Bininda-Emonds, M. Cardillo, J. Gittleman, K. Jones, C. Orme, and A. Purvis (2007). The fast-slow continuum in mammalian life history: An empirical reevaluation. *The American Naturalist 169*(6), 748–757.

Blackburn, T. (1991). Evidence for a 'fast-slow' continuum of life-history traits among parasitoid Hymenoptera. *Functional Ecology 5*(1), 65–74.

Blanquart, F., C. Wymant, M. Cornelissen, A. Gall, M. Bakker, D. Bezemer, M. Hall, M. Hillebregt, S. H. Ong, J. Albert, N. Bannert, J. Fellay, K. Fransen, A. J. Gourlay, M. K. Grabowski, B. Gunsenheimer-Bartmeyer, H. F. Günthard, P. Kivelä, R. Kouyos, O. Laeyendecker, K. Liitsola, L. Meyer, K. Porter, M. Ristola, A. van Sighem, G. Vanham, B. Berkhout, P. Kellam, P. Reiss, C. Fraser, and B. collaboration (2017, 06). Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in europe. *PLoS Biology 15*(6), 1–26.

Cantet, R. J., A. N. Birchmeier, and J. P. Steibel (2004). Full conjugate analysis of normal multiple traits with missing records using a generalized inverted Wishart distribution. *Genetics, Selection and Evolution 36*, 49–64.

Capellini, I., J. Baker, W. Allen, S. Street, and C. Vendetti (2015). The role of life history traits in mammalian invasion success. *Ecology Letters 18*, 1099–1107.

Clobert, J., T. Garland, and R. Barbault (1998). The evolution of demographic tactics in lizards: A test of some hypotheses concerning life history evolution. *Journal of Evolutionary Biology 11*(3), 329–364.

Cybis, G., J. Sinsheimer, T. Bedford, A. Mather, P. Lemey, and M. Suchard (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Annals of Applied Statistics 9*, 969 – 991.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society, series B 39*(1), 1–38.

Dill, K. A., K. Ghosh, and J. D. Schmit (2011). Physical limits of cells and proteomes. *Proceedings of the National Academy of Sciences 108*(44), 17876–17882.

Dominici, F., G. Parmigiani, and M. Clyde (2000). Conjugate analysis of multivariate normal data with incomplete observations. *Canadian Journal of Statistics 28*, 533–550.

Doolittle, W. F. and C. Sapienza (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature 284*(5757), 601.

Drummond, A. J., S. Y. Ho, M. J. Phillips, and A. Rambaut (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology 4*(5), e88.

Drummond, A. J., M. A. Suchard, and A. Rambaut (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution 29*, 1969–1973.

Ernest, S. (2003). Life history characteristics of placental nonvolant mammals. *Ecology 84*(12), 3402.

Falconer, D. S. (1960). *Introduction to Quantitative Genetics.* Oliver And Boyd; Edinburgh; London.

Felsenstein, J. (1985, January). Phylogenies and the comparative method. *The American Naturalist 125*(1), 1–15.

Freckleton, R. P. (2012). Fast likelihood calculations for comparative analyses. *Methods in Ecology and Evolution 3*(5), 940–947.

Fritz, S., O. Bininda-Edmonds, and A. Purvis (2009). Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecology Letters 12*(6), 538–549.

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

Giovannoni, S. J., J. C. Thrash, and B. Temperton (2014). Implications of streamlining theory for microbial ecology. *The ISME Journal 8*(8), 1553.

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

Goberna, M. and M. Verdú (2016). Predicting microbial traits with phylogenies. *The ISME Journal 10*(4), 959.

Ho, L. S. T. and C. Ané (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology 63*(3), 397–408.

Hodcroft, E., J. D. Hadfield, E. Fearnhill, A. Phillips, D. Dunn, S. O'Shea, D. Pillay, A. J. Leigh Brown, UK HIV Drug Resistance Database, and UK CHIC Study (2014). The contribution of viral genotype to plasma viral set-point in HIV infection. *PLoS Pathogens 10*(5), e1004112.

Hurst, L. D. and A. R. Merchant (2001). High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proceedings of the Royal Society of London B: Biological Sciences 268*(1466), 493–497.

Jeschke, J. and H. Kokko (2009). The roles of body size and phylogeny in fast and slow life histories. *Evolutionary Ecology 23*(6), 867–878.

Jones, K. E., J. Bielby, M. Cardillo, S. A. Fritz, J. O'Dell, C. Orme, K. Safi, W. Sechrest, E. H. Boakes, C. Carbone, C. Connolly, M. J. Cuttis, J. K. Foster, R. Grenyer, M. Habib, C. A. Plaster, S. A. Price, E. A. Rigby, J. Rist, A. Teacher, O. R. Bininda-Emonds, J. L. Gittleman, G. M. Mace, and A. Purvis (2009). PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology 90*(9), 2648.

Little, R. and D. Rubin (1987). *Statistical Analysis With Missing Data*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley.

Liu, J. S., W. H. Wong, and A. Kong (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 157–169.

Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, et al. (2004). ARB: a software environment for sequence data. *Nucleic Acids Research 32*(4), 1363–1371.

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences 104*(suppl 1), 8597–8604.

Mitov, V. and T. Stadler (2018). A practical guide to estimating the heritability of pathogen traits. *Molecular Biology and Evolution 35*(3), 756–772.

Musto, H., H. Naya, A. Zavala, H. Romero, F. Alvarez-Valín, and G. Bernadi (2004). Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS letters 573*(1-3), 73–77.

Orgel, L. E. and F. H. Crick (1980). Selfish DNA: the ultimate parasite. *Nature 284*(5757), 604.

Pagel, M. (1999, October). Inferring the historical patterns of biological evolution. *Nature 401*, 877–884.

Pruesse, E., J. Peplies, and F. O. Glöckner (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics 28*(14), 1823–1829.

Pybus, O. G., M. A. Suchard, P. Lemey, F. J. Bernardin, A. Rambaut, F. W. Crawford, R. R. Gray, N. Arinaminpathy, S. L. Stramer, M. P. Busch, and E. L. Delwart (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Procedings of the National Academy of Sciences 109*(37), 15066–15071.

Rambaut, A., T. T. Lam, L. Max Carvalho, and O. G. Pybus (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution 2*(1), vew007.

Réale, D., D. Garant, M. M. Humphries, P. Bergeron, V. Careau, and P.-O. Montiglio (2010). Personality and the emergence of the pace-of-life syndrome concept at the population level. *Philosophical Transactions of the Royal Society B: Biological Sciences 365*(1560), 4051–4063.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution 3*(2), 217–223.

Reynolds, J. (2003). Life histories and extinction risk. In T. Blackburn and K. Gaston (Eds.), *Macroecology: Concepts and Consequences*, pp. 195–217. Oxford: Blackwell Publishing Ltd.

Roff, D. A. (2002). *Life History Evolution*. Sunderland, Massachusetts. Sinauer Associates.

Sabath, N., E. Ferrada, A. Barve, and A. Wagner (2013). Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biology and Evolution 5*(5), 966–977.

Sæther, B.-E. and Ø. Bakke (2000). Avian life history variation and contribution of demographic traits to the population growth rate. *Ecology 81*(3), 642–653.

Salguro-Gómez, R. (2017). Applications of the fast-slow continuum and reproductive strategy framework of plant life histories. *New Phytologist 213*(3), 1618–1624.

Shuter, B. J., J. Thomas, W. D. Taylor, and A. M. Zimmerman (1983). Phenotypic correlates of genomic dna content in unicellular eukaryotes and other cells. *The American Naturalist 122*(1), 26–44.

Stearns, S. C. (1989). Trade-offs in life-history evolution. *Functional Ecology 3*(3), 259–268.

Suchard, M. A., P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution 4*(1), vey016.

Swiss HIV Cohort Study, F. Schoeni-Affolter, B. Ledergerber, M. Rickenbach, C. Rudin, H. F. Günthard, A. Telenti, H. Furrer, S. Yerly, and P. Francioli (2009). Cohort profile: the swiss hiv cohort study. *International Journal of Epidemiology 39*(5), 1179–1189.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences 17*(2), 57–86.

To, T.-H., M. Jung, S. Lycett, and O. Gascuel (2016, Jan). Fast dating using least-squares criteria and algorithms. *Systematic Biology 65*(1), 82–97.

Tolkoff, M. R., M. E. Alfaro, G. Baele, P. Lemey, and M. A. Suchard (2017). Phylogenetic factor analysis. *Systematic Biology 67*(3), 384–399.

Visscher, P. M., W. G. Hill, and N. R. Wray (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics 9*(4), 255.

Vrancken, B., P. Lemey, A. Rambaut, T. Bedford, B. Longdon, H. F. Günthard, and M. A. Suchard (2015). Simultaneously estimating evolutionary history and repeated traits phylogenetic signal: applications to viral and host phenotypic evolution. *Methods in Ecology and Evolution 6*, 67–82.

Wang, H.-C., E. Susko, and A. J. Roger (2006). On the correlation between genomic G + C content and optimal growth temperature in prokaryotes: data quality and confounding factors. *Biochemical and Biophysical Research Communications 342*(3), 681–684.

Wiedmann, M., R. Primicerio, A. Dolgov, C. Ottensen, and M. Aschan (2014). Life history variation in Barents Sea fish: Implications for sensitivity to fishing in a changing environment. *Ecology and Evolution 4*(18), 3596–3611.

Wilks, S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics 3*, 163–195.

Worobey, M., T. D. Watts, R. A. McKay, M. A. Suchard, T. Granade, D. E. Teuwen, B. A. Koblin, W. Heneine, P. Lemey, and H. W. Jaffe (2016, 11). 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature 539*(7627), 98–101.

Wu, H., Z. Zhang, S. Hu, and J. Yu (2012). On the molecular mechanism of GC content variation among eubacterial genomes. *Biology Direct 7*(1), 2.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution 39*(3), 306–314.

Detailed Calculations

SI 1. MATRIX INVERSION COMPUTATIONS

To evaluate the observed data likelihood, we must compute branch-deflated precisions $\mathbf{P}_i^\star = \left(\mathbf{P}_i^- + t_i \boldsymbol{\delta}_i \boldsymbol{\Sigma} \boldsymbol{\delta}_i\right)^-$ for $i = 1, \ldots, 2N - 2$. We demonstrate below that this matrix exists and is well-defined under the definition of our pseudo-inverse. Using the permutation matrix $\mathbf{C}_i$ from Section 2.1.2, we decompose the diffusion variance $\boldsymbol{\Sigma}$ and node precision $\mathbf{P}_i$ such that

$$\boldsymbol{\Sigma} = \mathbf{C}_i \begin{pmatrix} \boldsymbol{\Sigma}_i^{\mathrm{obs}} & \boldsymbol{\Sigma}_i^{\mathrm{ol}} & \boldsymbol{\Sigma}_i^{\mathrm{om}} \\ - & \boldsymbol{\Sigma}_i^{\mathrm{lat}} & \boldsymbol{\Sigma}_i^{\mathrm{lm}} \\ - & - & \boldsymbol{\Sigma}_i^{\mathrm{mis}} \end{pmatrix} \mathbf{C}_i^t \text{ and}$$

$$\mathbf{P}_i = \mathbf{C}_i \mathrm{diag}\left[\infty\mathbf{I}, \tilde{\mathbf{P}}_i, 0\mathbf{I}\right] \mathbf{C}_i^t,$$

for $i = 1, \ldots, 2N - 2$. We use this decomposition to identify that:

$$
\begin{aligned}
\mathbf{P}_i^\star &= \left(\mathbf{P}_i^- + t_i \boldsymbol{\delta}_i \boldsymbol{\Sigma} \boldsymbol{\delta}_i\right)^- \\
&= \mathbf{C}_i \left(\left(\mathrm{diag}\left[\infty\mathbf{I}, \tilde{\mathbf{P}}_i, 0\mathbf{I}\right]\right)^- + \mathrm{diag}\left[t_i \begin{pmatrix} \boldsymbol{\Sigma}_i^{\mathrm{obs}} & \boldsymbol{\Sigma}_i^{\mathrm{ol}} \\ - & \boldsymbol{\Sigma}_i^{\mathrm{lat}} \end{pmatrix}, 0\mathbf{I}\right]\right)^- \mathbf{C}_i^t \\
&= \mathbf{C}_i \left(\mathrm{diag}\left[0\mathbf{I}, \tilde{\mathbf{P}}_i^{-1}, \infty\mathbf{I}\right] + \mathrm{diag}\left[t_i \begin{pmatrix} \boldsymbol{\Sigma}_i^{\mathrm{obs}} & \boldsymbol{\Sigma}_i^{\mathrm{ol}} \\ - & \boldsymbol{\Sigma}_i^{\mathrm{lat}} \end{pmatrix}, 0\mathbf{I}\right]\right)^- \mathbf{C}_i^t \qquad (1) \\
&= \mathbf{C}_i \left(\mathrm{diag}\left[\mathbf{T}, \infty\mathbf{I}\right]\right)^- \mathbf{C}_i^t \\
&= \mathbf{C}_i \mathrm{diag}\left[\mathbf{T}^{-1}, 0\mathbf{I}\right] \mathbf{C}_i^t,
\end{aligned}
$$

where

$$\mathbf{T} = \mathrm{diag}\left[0\mathbf{I}, \tilde{\mathbf{P}}_i^{-1}\right] + t_i \begin{pmatrix} \boldsymbol{\Sigma}_i^{\mathrm{obs}} & \boldsymbol{\Sigma}_i^{\mathrm{ol}} \\ - & \boldsymbol{\Sigma}_i^{\mathrm{lat}} \end{pmatrix} = \begin{pmatrix} t_i \boldsymbol{\Sigma}_i^{\mathrm{obs}} & t_i \boldsymbol{\Sigma}_i^{\mathrm{ol}} \\ - & \tilde{\mathbf{P}}_i^{-1} + t_i \boldsymbol{\Sigma}_i^{\mathrm{lat}} \end{pmatrix}. \qquad (2)$$

The matrix $\mathbf{T}$ is the sum of a positive-definite matrix and positive-semidefinite matrix and is therefore invertible.

SI 2.   HERITABILITY STATISTIC

We compute the expectation of the empirical variance $\mathbb{E}[\mathbf{S}^2(\mathbf{Y})]$ under the MBD model with residual variance as follows:

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{S}^2(\mathbf{Y})\right] &= \mathbb{E}\left[\frac{1}{N}\left(\mathbf{Y}-\bar{\mathbf{Y}}\right)^t\left(\mathbf{Y}-\bar{\mathbf{Y}}\right)\right] \\
&= \frac{1}{N}\mathbb{E}\left[\mathbf{Y}^t\mathbf{Y}-\frac{2}{N}\mathbf{Y}^t\mathbf{J}_N\mathbf{Y}+\frac{1}{N^2}\mathbf{Y}^t\mathbf{J}_N\mathbf{J}_N\mathbf{Y}\right] \\
&= \frac{1}{N}\mathbb{E}\left[\mathbf{Y}^t\mathbf{Y}-\frac{2}{N}\mathbf{Y}^t\mathbf{J}_N\mathbf{Y}+\frac{1}{N}\mathbf{Y}^t\mathbf{J}_N\mathbf{Y}\right] \\
&= \frac{1}{N}\mathbb{E}\left[\mathbf{Y}^t\mathbf{Y}-\frac{1}{N}\mathbf{Y}^t\mathbf{J}_N\mathbf{Y}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathbf{Y}_i\mathbf{Y}_i^t\right]-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left[\mathbf{Y}_i\mathbf{Y}_j^t\right].
\end{aligned}
\tag{3}
$$

The multivariate normal distribution of $\mathrm{vec}\left[\mathbf{Y}\right]$ implies $\mathrm{Cov}\left(Y_{ik},Y_{jl}\right)=\Sigma_{kl}\Upsilon_{ij}+\Gamma_{kl}^{-1}1_{\{i=j\}}$ where $1_{\{i=j\}}$ is an indicator function. Using this information in SI Equation 3,

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{S}^2(\mathbf{Y})\right] &= \frac{1}{N}\sum_{i=1}^{N}\left(\Upsilon_{ii}\Sigma+\Gamma^{-1}+\mathbb{E}\left[\mathbf{Y}_i\right]\mathbb{E}\left[\mathbf{Y}_i\right]^t\right) \\
&\quad -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\Upsilon_{ij}\Sigma+\Gamma^{-1}1_{\{i=j\}}+\mathbb{E}\left[\mathbf{Y}_i\right]\mathbb{E}\left[\mathbf{Y}_j\right]^t\right) \\
&= \frac{1}{N}\mathrm{tr}\left[\Upsilon\right]\Sigma+\Gamma^{-1}-\left(\frac{1}{N^2}\mathbf{1}_N^t\Upsilon\mathbf{1}_N\right)\Sigma-\frac{1}{N}\Gamma^{-1} \\
&\quad +\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathbf{Y}_i\right]\mathbb{E}\left[\mathbf{Y}_i\right]^t-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left[\mathbf{Y}_i\right]\mathbb{E}\left[\mathbf{Y}_j\right]^t.
\end{aligned}
\tag{4}
$$

Note that $\mathbb{E}\left[\mathbf{Y}_i\right]=\mathbf{Y}_{2N-1}$ for $i=1\dots N$, which implies

$$
\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\mathbf{Y}_i\right]\mathbb{E}\left[\mathbf{Y}_i\right]^t-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left[\mathbf{Y}_i\right]\mathbb{E}\left[\mathbf{Y}_j\right]^t=0.
\tag{5}
$$

As such, our expression for the expected empirical variance reduces to the following:

$$
\mathbb{E}\left[\mathbf{S}^2(\mathbf{Y})\right]=\frac{N-1}{N}\Gamma^{-1}+\left(\frac{1}{N}\mathrm{tr}\left[\Upsilon\right]-\frac{1}{N^2}\mathbf{1}_N^t\Upsilon\mathbf{1}_N\right)\Sigma.
\tag{6}
$$

SI 3. ALGORITHM FOR EFFICIENTLY COMPUTING $\mathbf{H}$

Note that naive computation of $c_\sigma = \frac{1}{N}\mathrm{tr}\left[\boldsymbol{\Upsilon}\right] + \frac{1}{N^2}\mathbf{1}_N^t\boldsymbol{\Upsilon}\mathbf{1}_N$ in Equation 24 would require constructing the $N \times N$ matrix $\boldsymbol{\Upsilon}$ and summing over all its elements, which has computation complexity of at least $\mathcal{O}(N^2)$. For cases where $\mathscr{F}$ is random and changes throughout the MCMC simulation, this quantity must be re-computed each time we compute the statistic. To avoid this issue, we implement an algorithm that avoids constructing $\boldsymbol{\Upsilon}$ in its entirety and simply calculates both $\mathrm{tr}\left[\boldsymbol{\Upsilon}\right]$ and $\mathbf{1}_N^t\boldsymbol{\Upsilon}\mathbf{1}_N$ in $\mathcal{O}(N)$ time. The algorithm performs a post-order traversal of the tree where at each internal node ν_ℓ we compute $N_{\lfloor \ell \rfloor}$ (the number of tips below ν_ℓ), $s_{\lfloor \ell \rfloor}$ (the sum of all elements in $\boldsymbol{\Upsilon}_{\lfloor \ell \rfloor}$), and $d_{\lfloor \ell \rfloor}$ (the sum of the diagonal elements in $\boldsymbol{\Upsilon}_{\lfloor \ell \rfloor}$). We define $\boldsymbol{\Upsilon}_{\lfloor \ell \rfloor}$ as the tree variance-covariance matrix constructed from the sub-tree $\mathscr{F}_{\lfloor \ell \rfloor}$ that is simply the tree that contains only the nodes below ν_ℓ with node ν_ℓ as its root. For internal nodes ν_ℓ with child nodes ν_j and ν_k, we accumulate

$$
\begin{aligned}
N_{\lfloor \ell \rfloor} &= N_{\lfloor j \rfloor} + N_{\lfloor k \rfloor} + 1, \\
s_{\lfloor \ell \rfloor} &= s_{\lfloor j \rfloor} + s_{\lfloor k \rfloor} + t_j N_{\lfloor j \rfloor}^2 + t_k N_{\lfloor k \rfloor}^2, \text{ and} \\
d_{\lfloor \ell \rfloor} &= d_{\lfloor j \rfloor} + d_{\lfloor k \rfloor} + t_j N_{\lfloor j \rfloor} + t_k N_{\lfloor k \rfloor}.
\end{aligned}
\tag{7}
$$

At the tips, we initialize with $s_{\lfloor i \rfloor} = d_{\lfloor i \rfloor} = 0$ and $N_{\lfloor i \rfloor} = 1$. At the root, $s_{\lfloor 2N-1 \rfloor} = \mathbf{1}_N^t\boldsymbol{\Upsilon}\mathbf{1}_N$ and $d_{\lfloor 2N-1 \rfloor} = \mathrm{tr}\left[\boldsymbol{\Upsilon}\right]$. This algorithm visits each node in $\mathscr{F}$ exactly once and has run time $\mathcal{O}(N)$.