Paleobiology

www.cambridge.org/pab

Bayesian inference of phylogenetic trees is not misled by correlated discrete morphological characters

Xueer Liu[1,2] and Chi Zhang[1,2]

[1]Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China
[2]University of Chinese Academy of Sciences, Beijing 101408, China

Article

Cite this article: Liu, X., and C. Zhang (2025). Bayesian inference of phylogenetic trees is not misled by correlated discrete morphological characters. *Paleobiology*, 1–9.
https://doi.org/10.1017/pab.2025.10076

Received: 16 June 2025
Revised: 01 September 2025
Accepted: 03 September 2025

Handling Editor:
Rachel Warnock

Corresponding author:
Chi Zhang;
Email: zhangchi@ivpp.ac.cn

The authors contributed equally to this study.

Abstract

Morphological characters are central to phylogenetic inference, especially for fossil taxa for which genomic data are unavailable. While Bayesian methods have gained popularity in recent years, they typically assume characters evolve independently, despite known correlations among characters. Here, we assess the impact of character correlation and evolutionary rate heterogeneity on Bayesian phylogenetic inference using extensive simulations of binary characters evolving under independent and correlated models. We find that Bayesian inference assuming character independence accurately recovers tree topologies even when characters are strongly correlated or evolve under heterogeneous rates. However, branch lengths or clock rates tend to be underestimated, particularly under extreme rate heterogeneity. These biases are partially corrected using models that integrate over character-state heterogeneity. Our results demonstrate that Bayesian methods are robust to violations of character independence in topological inference, supporting their continued use in morphological phylogenetics.

Non-technical Summary

Scientists often use morphological traits to figure out how fossil species are related. A popular method to do this assumes each trait changes on its own, even though many traits can be linked. This study used computer simulations to see how much this assumption affects the results. The researchers found that even when traits are connected or change at different rates, the method still does a good job figuring out the species tree. However, it can make mistakes in estimating how fast the traits changed over time. Some improved models help fix these errors. Overall, the study shows that current methods work well for figuring out relationships among species using morphological traits.

Introduction

Phylogenetic inference is essential for answering various questions in evolutionary biology. Despite the tremendous amount of genomic data available, morphological characters remain the primary or sole information to infer phylogenies of fossil taxa and to study deep-time divergence (Lee and Palci 2015; Donoghue and Yang 2016). Discrete characters are the main type of data and are traditionally analyzed under maximum parsimony. In recent years, model-based methods, including maximum likelihood and Bayesian inference, have been shown to have comparable or better performance in inferring phylogenies (Wright and Hillis 2014; O'Reilly et al. 2016, 2018; Brown et al. 2017; Puttick et al. 2017, 2019; Smith 2019; Keating et al. 2020). Among these methods, characters are treated as independent features. The simplest model for discrete characters is the Mk model (Lewis 2001), in which the rates of changes among the k states are equal. The most frequently used Mkv model (with suffix "v"; Lewis 2001) is a variant that accounts for the ascertainment bias of coding only variable characters.

Correlation among characters has long been recognized. One classical example is the tail-presence and tail-color problem (Maddison 1993), because the two characters are logically dependent. Several algorithmic solutions have been proposed to handle such cases, under either parsimony- or model-based criteria (Brazeau et al. 2019; Goloboff et al. 2021; Hopkins and St. John 2021; Tarasov 2023). Another example is that some characters are functionally or developmentally dependent (Beaulieu and Donoghue 2013; Leslie et al. 2015; Billet and Bardin 2019). This study mainly deals with the latter but also covers the former if the inapplicable states are governed by a hidden process (Tarasov 2021). In general, correlated discrete characters can be modeled by a Markov chain with rates among states as parameters (Pagel 1994; Pagel and Meade 2006). However, such a model is typically employed in phylogenetic comparative methods to study the evolution of two or three characters given fixed trees (Pagel et al. 2004; Pagel and Meade 2006; Beaulieu and Donoghue 2013; Billet and Bardin 2019), but is rarely used for inference of

PALEOBIOLOGY A PUBLICATION OF THE Paleontological SOCIETY

CAMBRIDGE UNIVERSITY PRESS

phylogenetic trees, as the number of parameters grows so dramatically with the number of characters and the model quickly becomes unidentifiable. Instead, all inference methods (parsimony, maximum likelihood, and Bayesian) typically assume all the characters are independent (Felsenstein 1985a).

Simulation studies have shown that Bayesian inference assuming character independence outperforms parsimony-based solutions in the case of logical dependence (Simões et al. 2022). However, no study so far has investigated how the Bayesian method performs in the general case of character dependence. Previous simulations used the simplest Mkv model for inference (Wright and Hillis 2014; O'Reilly et al. 2016, 2018; Puttick et al. 2017, 2019; Smith 2019; Keating et al. 2020), thus rate heterogeneity in state changes and across characters was not considered. Those studies also focused on non-clock (unrooted) trees. Herein we perform computer simulations to study the performance of Bayesian inference assuming character independence, with data simulated under either independent or dependent evolution under various conditions of evolutionary rate heterogeneity. We perform both non-clock and tip-dating analyses, and in the latter case, fossil ages are used and the results are dated (rooted) timetrees (Pyron 2011; Ronquist et al. 2012a; Zhang et al. 2016).

Methods

Markov Models

In general, discrete character evolution can be modeled by a Markov chain with a Q-matrix specifying the rates of changes (Pagel 1994). We first describe the model for a single binary character, then the models for a doublet and a triplet of correlated binary characters. For simplicity, we do not further consider correlation of four or more characters, or characters with more than two states.

For a binary character, the changes between states 0 and 1 are determined by this instantaneous rate matrix

$$Q_1 = \lambda \begin{bmatrix} -\pi_1 & \pi_1 \\ \pi_0 & -\pi_0 \end{bmatrix},$$

and the transition probability matrix is

$$P(t) = \begin{bmatrix} \pi_0 + \pi_1 e^{-\lambda t} & \pi_1 - \pi_1 e^{-\lambda t} \\ \pi_0 - \pi_0 e^{-\lambda t} & \pi_1 + \pi_0 e^{-\lambda t} \end{bmatrix}.$$

This model extends the Mk model (Lewis [2001], in which $\pi_0 = \pi_1 = 0.5$), allowing the equilibrium state frequencies to vary (Ronquist and Huelsenbeck 2003; Klopfstein et al. 2015; Wright et al. 2016) and is a two-state variate of the F81 model (Felsenstein 1981). It has two free parameters (λ and π_0). The average rate of change is $2\lambda\pi_0\pi_1$. Because λ and t are multiplied together in the transition probability matrix, they are not identifiable without further assumptions about the time and/or the rate.

For a doublet of binary characters, the general model for the four state pairs, 00, 01, 10 and 11, is introduced with eight free parameters (Pagel 1994). The model is not necessarily time reversible, and the Q-matrix may have complex eigenvalues and eigenvectors. For mathematical convenience, we reparametrize the Q-matrix as

$$Q_2 = \begin{bmatrix} \cdot & a\pi_2 & b\pi_3 & 0 \\ a\pi_1 & \cdot & 0 & c\pi_4 \\ b\pi_1 & 0 & \cdot & d\pi_4 \\ 0 & c\pi_2 & d\pi_3 & \cdot \end{bmatrix} = \begin{bmatrix} \cdot & a & b & 0 \\ a & \cdot & 0 & c \\ b & 0 & \cdot & d \\ 0 & c & d & \cdot \end{bmatrix} \begin{bmatrix} \pi_1 & 0 & 0 & 0 \\ 0 & \pi_2 & 0 & 0 \\ 0 & 0 & \pi_3 & 0 \\ 0 & 0 & 0 & \pi_4 \end{bmatrix},$$

with $\{a, b, c, d\}$ as the exchangeability rates and $\boldsymbol{\pi} = \{\pi_1, \pi_2, \pi_3, \pi_4\}$ as the equilibrium state frequencies for the four state pairs. The model is then time-reversible with seven free parameters. This can be viewed as a special case of the GTR model (Tavaré 1986; Yang 1994a). Setting $q_{12} = q_{34}$, $q_{13} = q_{24}$, $q_{21} = q_{43}$, and $q_{31} = q_{42}$ results in independent evolution with four parameters ($\boldsymbol{\pi}$ is derived from $\{a, b, c, d\}$), and will be equivalent to the Mk model by further constraining $a = b = c = d$ (as few as one free parameter).

Similarly, we can use this rate matrix,

$$Q_3 = \begin{bmatrix} \cdot & a & b & 0 & i & 0 & 0 & 0 \\ a & \cdot & 0 & c & 0 & j & 0 & 0 \\ b & 0 & \cdot & d & 0 & 0 & k & 0 \\ 0 & c & d & \cdot & 0 & 0 & 0 & l \\ i & 0 & 0 & 0 & \cdot & e & f & 0 \\ 0 & j & 0 & 0 & e & \cdot & 0 & g \\ 0 & 0 & k & 0 & f & 0 & \cdot & h \\ 0 & 0 & 0 & l & 0 & g & h & \cdot \end{bmatrix} \begin{bmatrix} \pi_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \pi_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \pi_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \pi_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \pi_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \pi_6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \pi_7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \pi_8 \end{bmatrix},$$

for a triplet of binary characters with eight states, 000, 001, 010, 011, 100, 101, 110 and 111. Simultaneous changes of two or three states are negligible, so that their rates are zero. The model has 19 free parameters. The average rate is $-\sum_i \pi_i q_{ii}$, where q_{ii} is the i^{th} diagonal element in the Q-matrix.

Simulation Procedure

We first generated variable timetrees from a birth–death process using TreeSim in R (Stadler 2011) with a birth rate of 5.0 and a death rate of 4.0, conditioned on a root age of 1.0. The ages are on a relative scale and can be arbitrarily rescaled depending on the chosen time unit. From the trees simulated, we kept 100 trees with no more than 50 tips to make sure the data size is manageable. We simply treated the extinct tips as fossils, without further sampling fossils along the tree. The distribution of tree lengths and the numbers of extant and extinct tips are shown in Figure 1.

For each tree, we then simulated evolution of discrete morphological characters along the tree under various models and settings (Table 1). The general procedure was generating the exponential waiting times and using the jump chain given the Q-matrix (Yang 2014: section 12.5.4). This is particularly useful when the transition probability matrix is hard to derive. The starting state at the root was randomly drawn from the equilibrium frequencies ($\boldsymbol{\pi}$). Only variable characters were kept at the tips referring to empirical practices.

For independent binary characters, the simplest model is fixing $\pi_0 = \pi_1 = 0.5$ (referred to as M2v herein) for all characters, representing homogeneous evolution. To introduce heterogeneity in character states, we drew π_0 from a uniform distribution independently for each character and let $\pi_1 = 1 - \pi_0$ (the F81-alike extension, referred to as F2v herein). We rescaled Q_1 so that the average rate per character (i.e., the base clock rate) is 1.0, such that the branch lengths in the tree are measured by distance. To further introduce heterogeneity along time, each branch length for each character was multiplied by a relative rate r independently drawn from a lognormal distribution with mean 1.0 and variance 4.0, representing the most heterogeneous case (referred as "no common mechanism" [NCM]; Tuffley and Steel 1997). We recorded a moderate size of 200 variable characters in the data matrix in each replicate of the three settings.

We used the rate matrix Q_2 to simulate pairs of binary characters (referred as G4v herein). We drew $\{a, b, c, d\}$ and $\{\pi_1, \pi_2, \pi_3, \pi_4\}$ from a symmetric Dirichlet distribution with parameter 10 (representing slight correlation) or 1.0 (severe correlation) for each doublet. We rescaled Q_2 to have an average rate of 2.0 (per character rate being

https://doi.org/10.1017/pab.2025.10076 Published online by Cambridge University Press

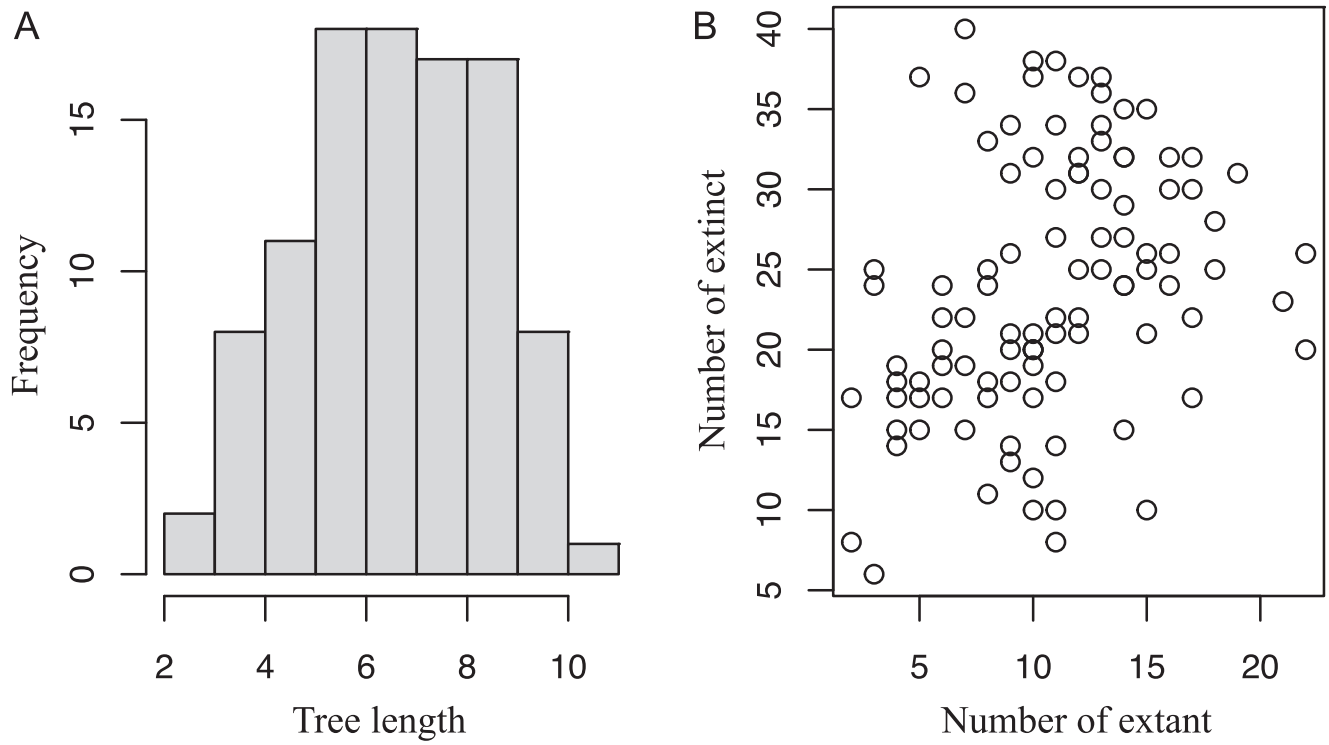

Figure 1. The distribution of tree length **(A)** and the numbers of extant and extinct tips **(B)** of the simulated trees.

Table 1. Models and settings used in the simulations and inferences. See "Methods" for the explanations of the symbols.

Simulation	Inference
M2v: r = 1.0, $\pi_0 = \pi_1 = 0.5$	M2v or F2v
F2v: r = 1.0, $\{\pi_0, \pi_1\}$ ~ SymDir(α = 1)	M2v or F2v
F2v: r ~ LNorm(m = 1, v = 4), $\{\pi_0, \pi_1\}$ ~ SymDir(1)	M2v or F2v+G
G4v: r = 1.0, $\{a,b,c,d\}$ and $\boldsymbol{\pi}$ ~ SymDir(α = 10)	M2v or F2v
G4v: r = 1.0, $\{a,b,c,d\}$ and $\boldsymbol{\pi}$ ~ SymDir(α = 1)	M2v or F2v
G4v: r ~ LNorm(m = 1, v = 4), $\{a,b,c,d\}$ and $\boldsymbol{\pi}$ ~ SymDir(α = 1)	M2v or F2v+G
G8v: r = 1.0, $\{a,\ldots,l\}$ and $\boldsymbol{\pi}$ ~ SymDir(α = 10)	M2v or F2v
G8v: r = 1.0, $\{a,\ldots,l\}$ and $\boldsymbol{\pi}$ ~ SymDir(α = 1)	M2v or F2v
G8v: r ~ LNorm(m = 1, v = 4), $\{a,\ldots,l\}$ and $\boldsymbol{\pi}$ ~ SymDir(α = 1)	M2v or F2v+G

1.0). To further have heterogeneous evolutionary rates, each branch length for each doublet is multiplied by an independent relative rate r as we did previously. We recorded 200 characters (i.e., 100 doublets) and ensured that all characters were variable in the data matrix.

Similarly, with three settings, we rescaled Q_3 to have an average rate of 3.0 (per character rate being 1.0), and simulated triplets of binary characters (referred as G8v). We recorded 201 variable binary characters (i.e., 67 triplets) but discarded the last character, so that we still had 200 characters in the data matrix. Considering that many empirical datasets are much smaller, we repeated the simulations under the same procedure with 50 variable correlated characters.

Phylogenetic Inference

Each data matrix was analyzed using the Bayesian phylogenetic inference software MrBayes 3.2.7 (Ronquist et al. 2012b). All characters were treated as independent, no matter how they were simulated. They were also treated as a single partition, meaning the branch lengths are shared by all characters (referred to as "common mechanism"; Tuffley and Steel 1997). This setting reflects the practice in most empirical analyses.

MrBayes supports both the M2v and F2v models. The M2v model has no free parameter other than the tree topology and branch lengths, while the F2v model has an extra parameter, π_0, which is averaged using a discretized symmetric beta prior with parameter α (Wright et al. 2016). We used an exponential hyperprior with mean 1.0 (Exp(1)) for α by default. For datasets simulated under NCM, we partially accommodated rate variation among characters using a discrete gamma distribution (Yang 1994b) (F2v+G; Table 1).

The non-clock analyses used the morphological data only and the branch lengths were measured by distance. As we simulated timetrees with both extant and extinct tips, we further incorporated the tip ages in another round of tip-dating analyses, so that we could disentangle the times and clock rates. The tip ages were assigned their true values assuming they are perfectly known. We specified diffuse Exp(1) prior for the root age and mean clock rate, used constant-rate fossilized birth–death prior (Stadler 2010) for the timetree and independent lognormal relaxed clock (Drummond et al. 2006) for the evolutionary rate variation, following common practices.

For each inference, two independent Markov chain Monte Carlo runs were executed each for 8 million generations with sampling frequency of 200. The beginning 35% of samples were discarded as burn-in, and the rest of the samples from the two runs were combined after checking consistency. We made sure the average

https://doi.org/10.1017/pab.2025.10076 Published online by Cambridge University Press

standard deviation of split frequencies was below 0.02, and the effective sample sizes were all greater than 100. In rare cases, we had to resume the analysis or double the chain length until these criteria were satisfied. The posterior tree samples were summarized as a 50% majority-rule consensus tree.

Missing Data

The main procedure involves no missing data. We also repeated the analyses with 50% missing states in the extinct taxa and 10% missing in the extant taxa, mimicking the observation in empirical datasets. Specifically, we replaced each state by a question mark in the data matrix with the corresponding probability (i.e., 0.1 for extant and 0.5 for extinct taxa). Such replacement was performed randomly on the generated binary characters rather than on the doublets or triplets.

Tree Distance Metrics

We employed both the Quartet (Estabrook et al. 1985) and Mutual Clustering Information (MCI; Smith 2020) metrics for comparing the inferred tree with the true tree generating the data. The MCI metric is a generalized Robinson-Foulds (RF) distance metric (Robinson and Foulds 1981) that is information based and less saturated; thus it is recommended over the RF metric (Smith 2020). The Quartet metric also has several advantages over the RF metric and is also recommended (Smith 2019).

Both distance metrics conflate accuracy and precision (Keating et al. 2020). Thus, we also calculated the Strict Joint Assertion (SJA, which is the number of quartets that are resolved identically in both trees over that resolved either identically or differently in both trees; Estabrook et al. 1985) as a measure of accuracy, and the percentage of resolved internal branches (the number of internal branches in the estimated consensus tree over that in the true tree) as precision.

The quartet-related metrics are calculated using the package Quartet in R (Smith 2019) and the MCI metric is using TreeDist in R (Smith 2020).

Results

We aim to investigate the performance of Bayesian phylogenetic inference using the M2v and F2v models by comparing the inferred tree with the true tree simulating the data. The Quartet and MCI metrics measure the topological differences, and the tree lengths in non-clock analyses and tree heights in tip-dating analyses represent the branch-length estimates.

We first look at the results from data without missing states. The first two scenarios represent rate homogeneous evolution, and the models used in the inference can match that in the simulation, in which M2v is a special case of F2v. They are the best-case scenarios and act as a baseline. The results do show that the topologies and branch lengths are inferred with good accuracy (Figs. 2A–D, 3A–D, cases 1, 2). For the following four scenarios, the rates in the *Q*-matrix are quite similar, as they were generated from a Dirichlet distribution with parameter 10. As a result, the performance of the Bayesian inference is almost the same as when there is no rate variation (Figs. 2A–D, 3A–D, cases 3–6).

The hardest situation appears to be when the data were simulated under F2v and each character has its own stationary frequencies (π_0 and π_1). The M2v model certainly does not account for this, resulting in larger tree distances (Fig. 2A–D, case 7) and underestimated tree lengths (Fig. 3A, case 7). In the tip-dating analyses, the tree height estimates are barely affected (Fig. 3B, case 7), but the clock rate is underestimated (Fig. 4A, case 7). The inference model of F2v is supposed to match the simulation condition; however, we did not estimate individual frequencies for each character due to identifiability issues. Instead, we averaged π_0 (and $\pi_1 = 1 - \pi_0$) using a discretized symmetric beta distribution (Wright et al. 2016). This strategy can correct the bias of tree-length or clock-rate estimates (Figs. 3A, 4A, case 8), but results in similar tree distances as using the M2v model (Fig. 2A–D, case 8). Having a further look at the accuracy (SJA) and precision (tree resolution) metrics, we find that the larger tree distances under M2v are largely contributed by decreased accuracy (Supplementary Figs. S1, S2, case 1 vs. case 7), whereas those under F2v are largely contributed by decreased precision (Supplementary Figs. S1, S2, case 2 vs. case 8).

Surprisingly, severe correlation in each pair or triplet of characters does not increase but instead decreases the tree distances (Fig. 2A–D, cases 9–12), although they still present higher distances than the homogeneous ones (Fig. 2A–D, cases 1–6). This results from both slightly increased accuracy and precision (Supplementary Figs. S1, S2, cases 7–12), likely because the rate heterogeneity for each character in these settings is slightly lower than that under independent evolution, which is reflected in the estimates of the shape parameter of the symmetric beta distribution (Supplementary Material, log files). Having more correlated characters (three vs. two) makes almost no difference in how the inference results are affected. Using the F2v model in inference cannot match the simulation models, but it is still helpful in slightly correcting the bias of underestimating tree length (Fig. 3A, cases 9–12) or clock rate (Fig. 4A, cases 9–12).

The most heterogeneous scenarios involve rate heterogeneity both among characters and across branches (NCM). However, using the simplest M2v model as well as the F2v model can achieve comparable performance as when there is no rate heterogeneity for inference of tree topology (Fig. 2A–D, cases 13–18, Supplementary Figs. S1, S2, cases 13–18). Evolutionary rate heterogeneity across branches appears to retain strong phylogenetic signal in the data. On the other hand, branch-length or clock-rate estimates are more biased, with M2v showing the most severe underestimation and narrowest credibility interval (CI) width (Figs. 3A, 4A, cases 13–18).

Empirical data typically contain many missing states. When 50% of states in the fossil taxa and 10% in the extant taxa are missing on average, we observe similar patterns as when there is no missing state, with decreased precision but similar accuracy (Figs. 2E–H, 3E–H, 4B, D, Supplementary Figs. S3, S4). In other words, missing data mostly result in more unresolved nodes in the trees and larger CIs of the tree lengths, but for the resolved part of the tree, the accuracy is similar to that when there is no missing data. Similar patterns are also observed when the number of characters is much smaller (50 vs. 200), with largely decreased precision and slightly decreased accuracy (Supplementary Fig. S5).

Discussion

The Bayesian method has been demonstrated to have good accuracy when data were generated under either common or no common mechanisms (Wright and Hillis 2014; O'Reilly et al. 2016, 2018; Puttick et al. 2017, 2019; Smith 2019; Keating et al. 2020). We moved one step further and introduced character correlation in the simulations. Both the non-clock and tip-dating analyses suggest

https://doi.org/10.1017/pab.2025.10076 Published online by Cambridge University Press

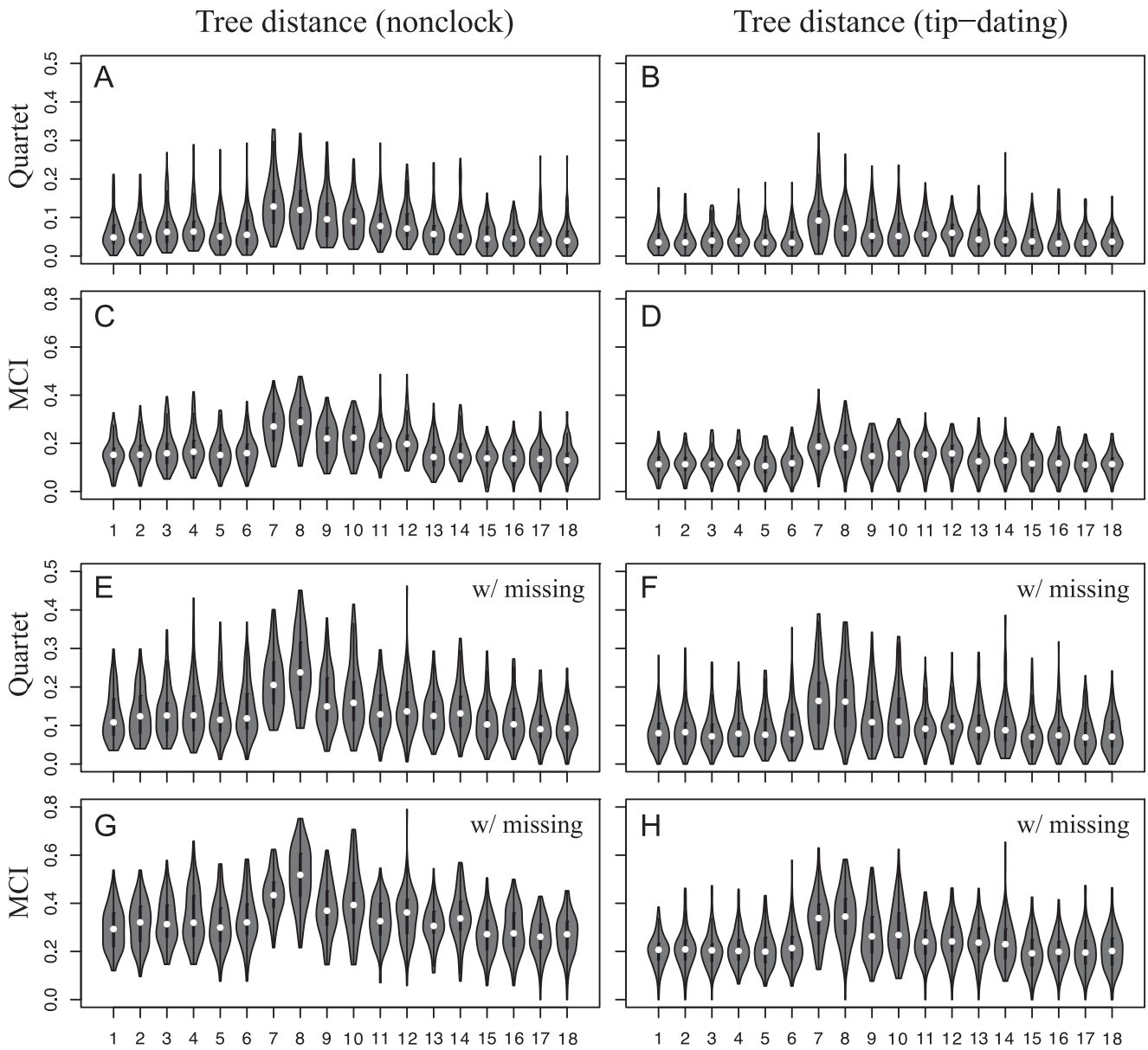

Figure 2. Tree distance metrics (Quartet and Mutual Clustering Information [MCI]) comparing the inferred tree with the true tree generating the data. Each violin plot contains 100 replicates. The left four panels show the results of non-clock analyses **(A, C, E, G)**, while the right four panels show the results of tip-dating analyses **(B, D, F, H)**. Panels labeled "w/ missing" **(E–H)** indicate scenarios with missing data. The numbers on the x-axis correspond to the following experiments (simulation model vs. inference model): 1, M2v–vs–M2v; 2, M2v–vs–F2v; 3, G4v(α = 10)–vs–M2v; 4, G4v(α = 10)–vs–F2v; 5, G8v(α = 10)–vs–M2v; 6, G8v(α = 10)–vs–F2v; 7, F2v(α = 1)–vs–M2v; 8, F2v(α = 1)–vs–F2v; 9, G4v(α = 1)–vs–M2v; 10, G4v(α = 1)–vs–F2v; 11, G8v(α = 1)–vs–M2v; 12, G8v(α = 1)–vs–F2v; 13, F2v(α = 1, *v* = 4)–vs–M2v; 14, F2v(α = 1, *v* = 4)–vs–F2v; 15, G4v(α = 1, *v* = 4)–vs–M2v; 16, G4v(α = 1, *v* = 4)–vs–F2v; 17, G8v(α = 1, *v* = 4)–vs–M2v; 18, G8v(α = 1, *v* = 4)–vs–F2v.

that Bayesian inference assuming character independence does not mislead the inference of tree topology when character correlation and rate heterogeneity are present. This is quite reassuring, as correlation and NCM have been argued to be quite common in morphological characters, and model-based methods are blamed for not accounting for these (Goloboff et al. 2018, 2019).

However, when the interest is the branch lengths, they can be biased toward underestimation when evolutionary rate variation is high among characters and along branches. Such variation can be modeled by general Markov processes in theory, but they are typically not practical in inference. Unlike molecular sequences where the same nucleotide across sites has the same biological meaning, morphological characters coded as 0, for example, have different meanings among characters; thus using one parameter for all the 0s would be pointless, whereas unlinking all of them would result in too many parameters. The best strategy so far has been using the F2v model, in which the state frequencies are averaged analogous to averaging the site rates (Wright et al. 2016). According to our simulation results, it is recommended over the M2v model in all the scenarios we have tested. However, the F2v model only accounts for rate variation among character states. To further account for rate variation along branches, we could subdivide the data into multiple partitions (e.g., according to the anatomical regions) and infer independent evolutionary rates for each partition (e.g., using unlinked clock models; Lee 2016; Zhang and Wang 2019). Bear in mind, though, we should keep enough (probably at

https://doi.org/10.1017/pab.2025.10076 Published online by Cambridge University Press

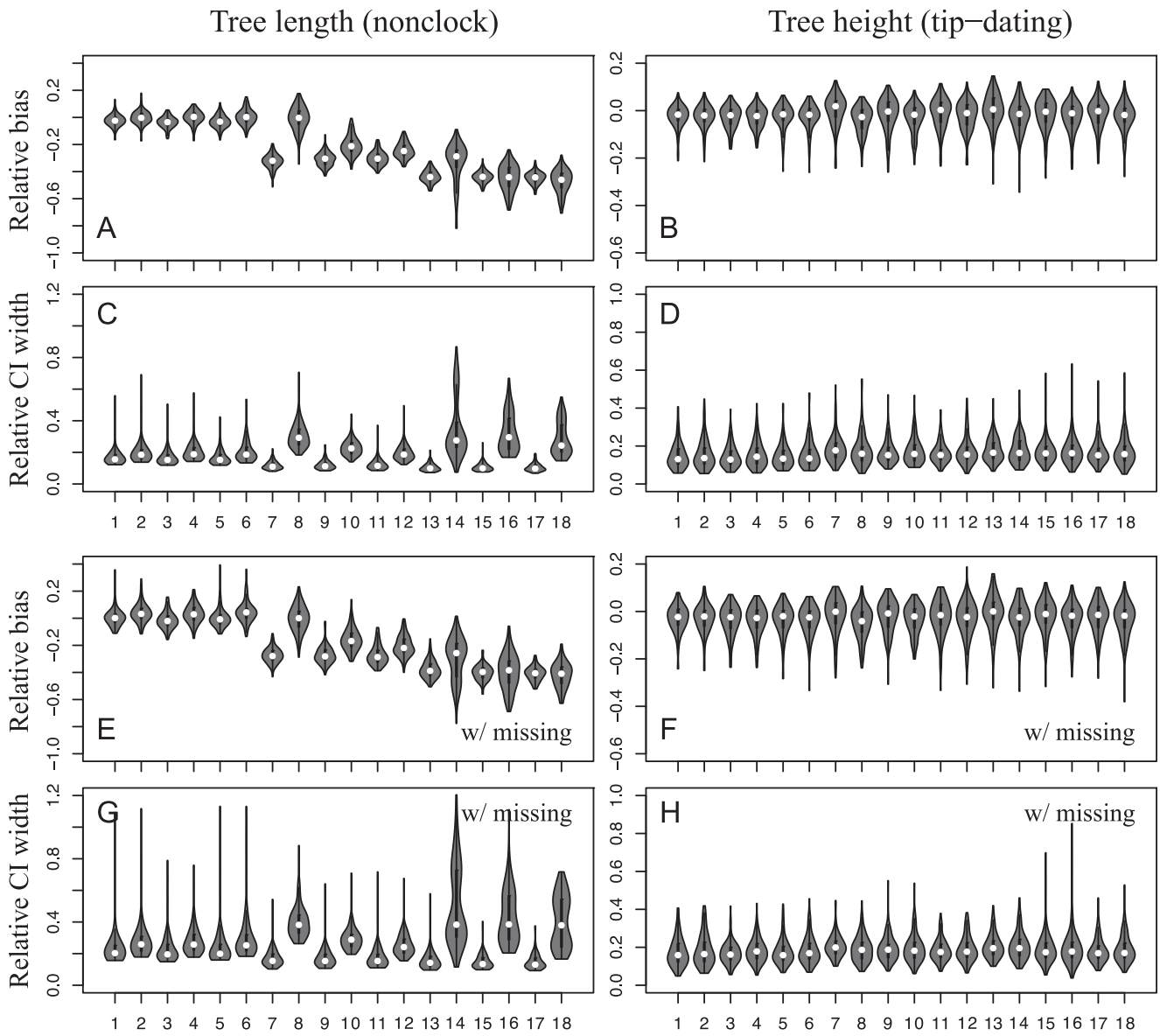

Figure 3. Relative bias (posterior mean minus the true value, then divided by the true value) and relative width of credibility interval (CI) (95% CI width divided by the true value) for each of the following experiments (simulation model vs. inference model): 1, M2v–vs–M2v; 2, M2v–vs–F2v; 3, G4v(α = 10)–vs–M2v; 4, G4v(α = 10)–vs–F2v; 5, G8v(α = 10)–vs–M2v; 6, G8v(α = 10)–vs–F2v; 7, F2v(α = 1)–vs–M2v; 8, F2v(α = 1)–vs–F2v; 9, G4v(α = 1)–vs–M2v; 10, G4v(α = 1)–vs–F2v; 11, G8v(α = 1)–vs–M2v; 12, G8v(α = 1)–vs–F2v; 13, F2v(α = 1, *v* = 4)–vs–M2v; 14, F2v(α = 1, *v* = 4)–vs–F2v; 15, G4v(α = 1, *v* = 4)–vs–M2v; 16, G4v(α = 1, *v* = 4)–vs–F2v; 17, G8v(α = 1, *v* = 4)–vs–M2v; 18, G8v(α = 1, *v* = 4)–vs–F2v. Each violin plot contains 100 replicates. The left four panels show the tree lengths from non-clock analyses **(A, C, E, G)**, while the right four panels show the tree heights from tip-dating analyses **(B, D, F, H)**. Panels labeled "w/ missing" (E–H) indicate scenarios with missing data.

least dozens of) characters in each partition to avoid overparameterization, especially when the data contain a large portion of missing states. Alternatively, a new method has been developed to account for rate variation both across characters and along branches by switching the rates among different rate regimes (Khakurel and Höhna 2025).

In the tip-dating analyses, we fixed the fossil ages to their true values. Hence the inferred tree heights (root ages) are reliable in all different conditions. This implies that incorporating accurate fossil information is crucial for dating divergence times, even when the morphological evolutionary model is mis-specified (Klopfstein et al. 2019). In practice, however, uncertainties in fossil ages and prior for the timetree (root age in particular) are likely to decrease the accuracy (Barido-Sottani et al. 2019; Luo et al. 2019). Depending on the data and models, the situation can become rather complicated (Simões et al. 2020; May et al. 2021). Optimistically, when the timetree is presumably reliable, evolutionary rate estimates could be refined using subsequent comparative methods for the characters of interest under more complex models (Pennell et al. 2014; Revell 2024).

We only considered correlated discrete morphological characters in this study. It is worth noting that there is a large body of literature for correlated continuous traits. The evolution of the traits is typically modeled by a Brownian motion (BM) (Felsenstein 1973, 1985b; Freckleton 2012) or an Ornstein–Uhlenbeck (OU) process (Uhlenbeck and Ornstein 1930; Felsenstein 1988; Hansen 1997; Butler and King 2004), and trait correlations are described by the variance–covariance matrix in the model. Relative to this, the

https://doi.org/10.1017/pab.2025.10076 Published online by Cambridge University Press

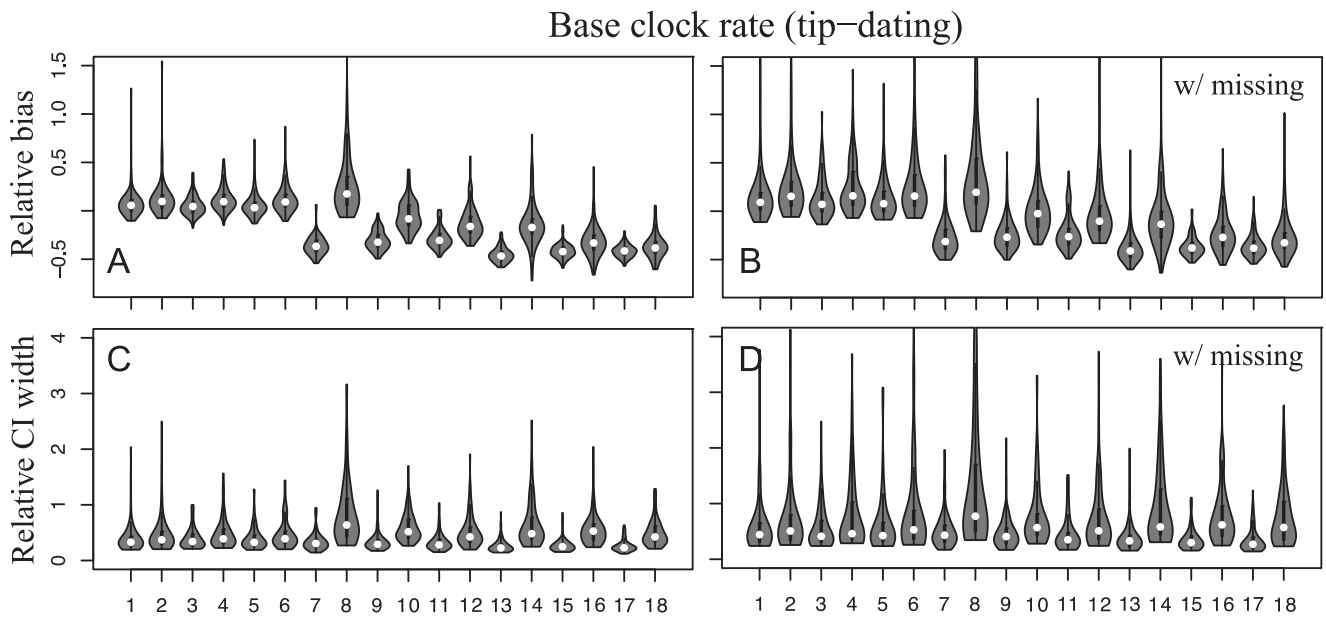

Figure 4. Relative bias (posterior mean minus the true value, then divided by the true value) and relative width of credibility interval (CI) (95% CI width divided by the true value) of the base clock rate for each of the following experiments (simulation model vs. inference model): 1, M2v–vs–M2v; 2, M2v–vs–F2v; 3, G4v(α = 10)–vs–M2v; 4, G4v(α = 10)–vs–F2v; 5, G8v(α = 10)–vs–M2v; 6, G8v(α = 10)–vs–F2v; 7, F2v(α = 1)–vs–M2v; 8, F2v(α = 1)–vs–F2v; 9, G4v(α = 1)–vs–M2v; 10, G4v(α = 1)–vs–F2v; 11, G8v(α = 1)–vs–M2v; 12, G8v(α = 1)–vs–F2v; 13, F2v(α = 1, *v* = 4)–vs–M2v; 14, F2v(α = 1, *v* = 4)–vs–F2v; 15, G4v(α = 1, *v* = 4)–vs–M2v; 16, G4v(α = 1, *v* = 4)–vs–F2v; 17, G8v(α = 1, *v* = 4)–vs–M2v; 18, G8v(α = 1, *v* = 4)–vs–F2v. Each violin plot contains 100 replicates. The left two panels are scenarios without missing data **(A and C)**, while the right two panels labeled with "w/ missing" **(B and D)** indicate scenarios with missing data.

threshold model (Wright 1934; Felsenstein 2005) is a promising alternative for correlated discrete characters, in which the observed discrete states depend on whether the underlying continuous trait (called liability) is above a threshold value. Although the BM and OU models have been well studied mathematically, practical implementations for phylogenetic inference are sparse (Álvarez-Carretero et al. 2019; Hassler et al. 2022; Zhang et al. 2023). The main reason is that these models are parameter-rich, and developing efficient computational methods is technically challenging. Thus, it appears to be an important area for further improvement.

Conclusion

Our results demonstrate that Bayesian inference of phylogenetic trees is remarkably robust to violations of the character independence assumption. Topological inference remains accurate across a range of realistic evolutionary scenarios, including strong correlation and substantial evolutionary rate heterogeneity among morphological characters. However, our analyses also reveal that branch lengths or clock rates may be systematically underestimated under simpler models when rate variation is present. To mitigate this, we recommend using models that average over character-state frequencies (e.g., F2v) and, when feasible, incorporating rate variation across partitions.

While this study focuses on discrete morphological characters, future work should extend to continuous trait and threshold models that more directly account for trait correlations. Overall, our findings support the continued and expanded use of Bayesian methods in morphological phylogenetics and call for methodological innovations to improve branch-length estimation under complex evolutionary processes.

Acknowledgments. This research was supported by the National Key Research and Development Program of China (2023YFF0804502 to C.Z.) and the National Natural Science Foundation of China (42172006 to C.Z.).

Author Contribution. C.Z. designed the study and led the analyses. X.L. performed the simulations. C.Z. led the interpretation of the results and writing of the manuscript. X.L. and C.Z. revised the manuscript and gave final approval for publication.

Competing Interests. The authors declare no competing interests.

Data Availability Statement. Electronic Supplementary Material is available from the GitHub Digital Repository: https://github.com/zhangchicool/morphSim.

Literature Cited

Álvarez-Carretero, S., A. Goswami, Z. Yang, and M. dos Reis. 2019. Bayesian estimation of species divergence times using correlated quantitative characters. *Systematic Biology* **68**:967–986.

Barido-Sottani, J., G. Aguirre-Fernández, M. J. Hopkins, T. Stadler, and R. Warnock. 2019. Ignoring stratigraphic age uncertainty leads to erroneous estimates of species divergence times under the fossilized birth-death process. *Proceedings of the Royal Society B* **286**:20190685.

Beaulieu, J. M., and M. J. Donoghue. 2013. Fruit evolution and diversification in campanulid angiosperms. *Evolution* **67**:3132–3144.

Billet, G., and J. Bardin. 2019. Serial homology and correlated characters in morphological phylogenetics: modeling the evolution of dental crests in placentals. *Systematic Biology* **68**:267–280.

Brazeau, M. D., T. Guillerme, and M. R. Smith. 2019. An algorithm for morphological phylogenetic analysis with inapplicable data. *Systematic Biology* **68**:619–631.

Brown, J. W., C. Parins-Fukuchi, G. W. Stull, O. M. Vargas, and S. A. Smith. 2017. Bayesian and likelihood phylogenetic reconstructions of morphological traits are not discordant when taking uncertainty into consideration: a comment on Puttick et al. *Proceedings of the Royal Society B* **284**:20170986.

https://doi.org/10.1017/pab.2025.10076 Published online by Cambridge University Press

Butler, M. A., and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *American Naturalist* **164**:683–695.

Donoghue, P. C. J., and Z. Yang. 2016. The evolution of methods for establishing evolutionary timescales. *Philosophical Transactions of the Royal Society B* **371**:20160020.

Drummond, A., S. Ho, M. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**:e88.

Estabrook, G. F., F. R. McMorris, and C. A. Meacham. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology* **34**:193–200.

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics* **25**:471–492.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**:368–376.

Felsenstein, J. 1985a. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**:783–791.

Felsenstein, J. 1985b. Phylogenies and the comparative method. *American Naturalist* **125**:1–15.

Felsenstein, J. 1988. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics* **19**:445–471.

Felsenstein, J. 2005. Using the quantitative genetic threshold model for inferences between and within species. *Philosophical Transactions of the Royal Society B* **360**:1427–1434.

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

Goloboff, P. A., A. Torres, and J. S. Arias. 2018. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics* **34**:407–437.

Goloboff, P. A., M. Pittman, D. Pol, and X. Xu. 2019. Morphological data sets fit a common mechanism much more poorly than DNA sequences and call into question the Mkv model. *Systematic Biology* **68**:494–504.

Goloboff, P. A., J. D. Laet, D. Ríos-Tamayo, and C. A. Szumik. 2021. A reconsideration of inapplicable characters, and an approximation with step-matrix recoding. *Cladistics* **37**:596–629.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* **51**:1341–1351.

Hassler, G. W., A. F. Magee, Z. Zhang, G. Baele, P. Lemey, X. Ji, M. Fourment, and M. A. Suchard. 2022. Data integration in Bayesian phylogenetics. *Annual Review of Statistics and Its Application* **10**:353–377.

Hopkins, M. J., and K. St. John. 2021. Incorporating hierarchical characters into phylogenetic analysis. *Systematic Biology* **70**:1163–1180.

Keating, J. N., R. S. Sansom, M. D. Sutton, C. G. Knight, and R. J. Garwood. 2020. Morphological phylogenetics evaluated using novel evolutionary simulations. *Systematic Biology* **69**:897–912.

Khakurel, B., and S. Höhna. 2025. A covarion model for phylogenetic estimation using discrete morphological datasets. *bioRxiv* 660793.

Klopfstein, S., L. Vilhelmsen, and F. Ronquist. 2015. A nonstationary Markov model detects directional evolution in hymenopteran morphology. *Systematic Biology* **64**:1089–1103.

Klopfstein, S., R. Ryer, M. Coiro, and T. Spasojevic. 2019. Mismatch of the morphology model is mostly unproblematic in total-evidence dating: insights from an extensive simulation study. *bioRxiv* 679084.

Lee, M. S. Y. 2016. Multiple morphological clocks and total-evidence tip-dating in mammals. *Biology Letters* **12**:20160033.

Lee, M. S. Y., and A. Palci. 2015. Morphological phylogenetics in the genomic age. *Current Biology* **25**:R922–R929.

Leslie, A. B., J. M. Beaulieu, P. R. Crane, P. Knopf, and M. J. Donoghue. 2015. Integration and macroevolutionary patterns in the pollination biology of conifers. *Evolution* **69**:1573–1583.

Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology* **50**:913–925.

Luo, A., D. A. Duchêne, C. Zhang, C.-D. Zhu, and S. Y. W. Ho. 2019. A simulation-based evaluation of tip-dating under the fossilized birth–death process. *Systematic Biology* **69**:325–344.

Maddison, W. P. 1993. Missing data versus missing characters in phylogenetic analysis. *Systematic Biology* **42**:576–581.

May, M. R., D. L. Contreras, M. A. Sundue, N. S. Nagalingum, C. V. Looy, and C. J. Rothfels. 2021. Inferring the total-evidence timescale of marattialean fern evolution in the face of model sensitivity. *Systematic Biology* **70**:1232–1255.

O'Reilly, J. E., M. N. Puttick, L. Parry, A. R. Tanner, J. E. Tarver, J. Fleming, D. Pisani, and P. C. J. Donoghue. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biology Letters* **12**:20160081.

O'Reilly, J. E., M. N. Puttick, D. Pisani, and P. C. J. Donoghue. 2018. Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. *Palaeontology* **61**:105–118.

Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B* **255**:37–45.

Pagel, M., and A. Meade. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *American Naturalist* **167**:808–825.

Pagel, M., A. Meade, and D. Barker. 2004. Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology* **53**:673–684.

Pennell, M. W., J. M. Eastman, G. J. Slater, J. W. Brown, J. C. Uyeda, R. G. FitzJohn, M. E. Alfaro, and L. J. Harmon. 2014. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* **30**:2216–2218.

Puttick, M. N., J. E. O'Reilly, A. R. Tanner, J. F. Fleming, J. Clark, L. Holloway, J. Lozano-Fernandez, et al. 2017. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proceedings of the Royal Society B* **284**:20162290.

Puttick, M. N., J. E. O'Reilly, D. Pisani, and P. C. J. Donoghue. 2019. Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without a probabilistic model. *Palaeontology* **62**:1–17.

Pyron, R. A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. *Systematic Biology* **60**:466–481.

Revell, L. J. 2024. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ* **12**:e16505.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* **53**:131–147.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**:1572–1574.

Ronquist, F., S. Klopfstein, L. Vilhelmsen, S. Schulmeister, D. L. Murray, and A. P. Rasnitsyn. 2012a. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Systematic Biology* **61**:973–999.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012b. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**:539–542.

Simões, T. R., M. W. Caldwell, and S. E. Pierce. 2020. Sphenodontian phylogeny and the impact of model choice in Bayesian morphological clock estimates of divergence times and evolutionary rates. *BMC Biology* **18**:191.

Simões, T. R., O. V. Vernygora, B. A. S. de Medeiros, and A. M. Wright. 2022. Handling logical character dependency in phylogenetic inference: extensive performance testing of assumptions and solutions using simulated and empirical data. *Systematic Biology* **72**:662–680.

Smith, M. R. 2019. Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. *Biology Letters* **15**:20180632.

Smith, M. R. 2020. Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. *Bioinformatics* **36**:5007–5013.

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

Stadler, T. 2011. Simulating trees with a fixed number of extant species. *Systematic Biology* **60**:676–684.

Tarasov, S. 2021. Integration of anatomy ontologies and evo-devo using structured Markov models suggests a new framework for modeling discrete phenotypic traits. *Systematic Biology* **68**:698–716.

Tarasov, S. 2023. New phylogenetic Markov models for inapplicable morphological characters. *Systematic Biology* **72**:681–693.

https://doi.org/10.1017/pab.2025.10076 Published online by Cambridge University Press

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

Tuffley, C., and M. Steel. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology* **59**:581–607.

Uhlenbeck, G. E., and L. S. Ornstein. 1930. On the theory of the Brownian motion. *Physical Review* **36**:823–841.

Wright, A. M., and D. M. Hillis. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE* **9**:e109210.

Wright, A. M., **G. T. Lloyd, and D. M. Hillis**. 2016. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology* **65**:602–611.

Wright, S. 1934. An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics* **19**:506.

Yang, Z. 1994a. Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution* **39**:105–111.

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

Yang, Z. 2014. *Molecular evolution: a statistical approach*. Oxford University Press, Oxford.

Zhang, C., and M. Wang. 2019. Bayesian tip dating reveals heterogeneous morphological clocks in Mesozoic birds. *Royal Society Open Science* **6**:182062.

Zhang, C., **T. Stadler**, **S. Klopfstein**, **T. Heath, and F. Ronquist**. 2016. Total-evidence dating under the fossilized birth-death process. *Systematic Biology* **65**:228–249.

Zhang, R., **A. J. Drummond, and F. K. Mendes**. 2023. Fast Bayesian Inference of Phylogenies from Multiple Continuous Characters. *Systematic Biology* **73**: 102–124.

https://doi.org/10.1017/pab.2025.10076 Published online by Cambridge University Press