Supplemental Text for: Mistreating birth-death models as priors in phylogenetic analysis compromises our ability to compare models

Contents

S§1 Simulation Study

To demonstrate the issues that arise from not treating samples as data, we performed a series of experiments in `RevBayes` (Höhna et al. 2016), which we selected for convenience; the problems we demonstrate are inherent to the standard treatment of tree models as priors and are independent of any particular implementation of these models. For each of the three experiments below, we performed a series of simulations under a specific model, and then compared the fit of a pair of competing models to these simulated data, using Bayes factors calculated with marginal-likelihood- and posterior-probability-based approaches. For the marginal-likelihood-based approach, we computed Bayes factors using a power-posterior algorithm, stepping-stone MCMC (SS MCMC; Xie et al. 2011), and for the posterior-probability-based approach we used reversible-jump MCMC (RJ MCMC; Green 1995). In all cases, we fixed the tree topology to the true value for computational tractability, but estimated the node ages, the birth-death parameters, and the parameters that governed the process of character evolution; for fossilized birth-death datasets, we also estimated whether each fossil was a sampled ancestor (*i.e.*, whether it was sampled along a branch leading to another sample).

Initial experiments indicated that sufficiently precise marginal-likelihood estimates using SS MCMC would be computationally prohibitive for large trees and sequence datasets. We therefore simulated relatively small trees and character datasets (exact sizes described below). Further, we implemented an adaptive power-posterior algorithm in `RevBayes`, similar to the one proposed by Friel et al. (2014). Power-posterior algorithms work by running a set of Markov chains, each with a "power", β_i, ranging between 0 and 1. For any given β_i (a "stone"), the chain samples from the distorted posterior distribution:

$$P_i(\theta \mid X) \propto P_i(X \mid \theta)^{\beta_i} P(\theta),$$

so that $\beta = 1$ corresponds to sampling from the posterior, and $\beta = 0$ corresponds to sampling from the prior. The sampled likelihood values among the separate stones—$P_i(X \mid \theta)$—can then be used to estimate the marginal likelihood, *e.g.*, using the stepping-stone estimator (Xie et al. 2011). Usually, the number of stones and the values of β are fixed in advance, but in our analyses we found that accurate marginal-likelihood estimates demanded a large number of stones, so adopted an adaptive approach. Briefly, our adaptive algorithm begins with two stones, $\beta_1 = 1$ and $\beta_2 = 0$, and then places additional stones until the estimate of the marginal likelihood converges; as with the original algorithm, the number of MCMC samples per stone is fixed in advance.

For each analysis described below, we performed two replicates to ensure stability of marginal-likelihood and posterior-ratio estimates. Our simulated data and code (including specific parameter settings for simulations and analyses) are available at Zenodo (`http://doi.org/10.5281/zenodo.5072533`) and GitHub (`https://github.com/mikeryanmay/bd_bayes_factors/releases/tag/initial_submission`).

S§1.1 Molecular substitution models

To demonstrate that SS and RJ MCMC compute the same BFs (and also to demonstrate that both of these methods are implemented correctly, *i.e.*, that our results are not a consequence of programming errors), we compared the fit of competing substitution models to simulated molecular datasets. We simulated ten trees under a birth-death (BD) model for each of four numbers of extant samples, $n = \{8, 16, 32, 64\}$. We assumed the tree began with two species at time $t = 1$, diversified at rates $\lambda = 4$ and $\mu = 2$, and that all extant species were sampled. For each tree, we simulated a nucleotide dataset with 100 sites under a Jukes-Cantor (JC; Jukes and Cantor 1969) model with rate parameter r

(scaled such that the expected number of substitutions per site was three). For each of the 40 simulated datasets, we computed Bayes factors between JC69 and K80 (Kimura 1980) substitution models using SS and RJ MCMC as described above, assuming the tree evolved under the true birth-death model (Fig. 2A, main text). Positive values of 2 ln BF indicate support for the variable-rate model.

S§1.2 Contemporaneous birth-death processes

Our second experiment considers the case of comparing two birth-death models for extant (contemporaneous) samples. We analyzed the same datasets simulated in the previous section, but in this case compared two tree models. The first model, M_1, is the Yule model (with speciation rate λ, and no extinction rate parameter), and the second model, M_2, is a BD model (with speciation rate λ and extinction rate μ); M_2 is the same tree model used to simulate the data, as described above. We computed Bayes factors between Yule (M_1) and BD (M_2) processes, again using both SS and RJ MCMC, assuming the sequence data evolved under the true JC69 substitution model (Fig. 2B, main text). In this case, positive values of 2 ln BF indicate support for the true model.

S§1.3 Fossilized birth-death processes

Our third experiment considers the more complex case of comparing two birth-death models for non-contemporaneous samples. We simulated four fossilized birth-death trees under a model that allowed the fossilization rate to vary. Specifically, each tree began with one lineage at time $t = 1$ (in the past, with $t = 0$ the present) and initially evolved under a fossilized birth-death model with $\lambda = 4$, $\mu = 2$, $\phi = 3$; at time $t = 0.5$ in the past, the fossilization rate changed to $\phi = 0.5$ (*i.e.*, the fossilization rate was high in the early part of the process, and decreased in the second half by a factor of six). We then simulated stratigraphic uncertainty by dividing time into 20 equally sized bins and using the boundaries of the bin that a given fossil sample fell into as the minimum and maximum ages of the sample (we assigned extant samples minimum and maximum ages of zero). For each tree, we simulated 100 binary characters under an Mk model (Lewis 2001) with rate r (scaled such that the expected number of substitutions per site was three). We then compared the fit of two competing fossilized birth-death models to the simulated data. The first model, M_1, has constant speciation, extinction, and fossilization rates (λ, μ, and ϕ, respectively). The second model, M_2, is the same as M_1, but allows the fossilization rate to vary over time. Specifically, the initial fossilization rate (at time $t = 1$ in the past) is ϕ_1, and at time $t = 0.5$ units in the past, it changes to rate ϕ_2, which persists until the present ($t = 0$). M_2 is similar to the simulating process in that the fossilization rate is not constant, and the time of the rate change is fixed; however, for both M_1 and M_2, we assume that the speciation, extinction, and fossilization rates are unknown. For both models, we also assume that all extant species are sampled, $\rho = 1$. We computed the Bayes factors between M_1 and M_2 using SS and RJ MCMC, assuming the morphological data evolves under the Mk model (Fig. 2C, main text). Positive values of 2 ln BF indicate support for the variable-rate model.

S§1.4 Extended Results

Our results indicate that SS and RJ MCMC provide essentially identical Bayes factors when comparing models of molecular evolution (simulation 1, Fig. 1A, main text), as expected based on the theoretical equivalence of these estimators. However, these method produce disparate estimates when comparing tree models, either for contemporaneous lineages (simulation 2, Fig. 1B, main text) or when including non-contemporaneous lineages (simulation 3, Fig. 1C, main text).

The discrepancy is not a consequence of a programming bug: the comparison of substitution models demonstrates that both algorithms are correctly implemented. Likewise, it is not a consequence of numerical MCMC errors: we performed two replicates of each of the analysis to confirm that Bayes-factor estimates were sufficiently precise both for SS-based estimates (Fig. S1) and RJ-based estimates (Fig. S2). For the contemporaneous birth-death models, we can make precise quantitative predictions about the magnitude of the discrepancy (Fig. 1B, colored dashed lines, main text), as we explain in Section S§4; our simulation results match these predictions.

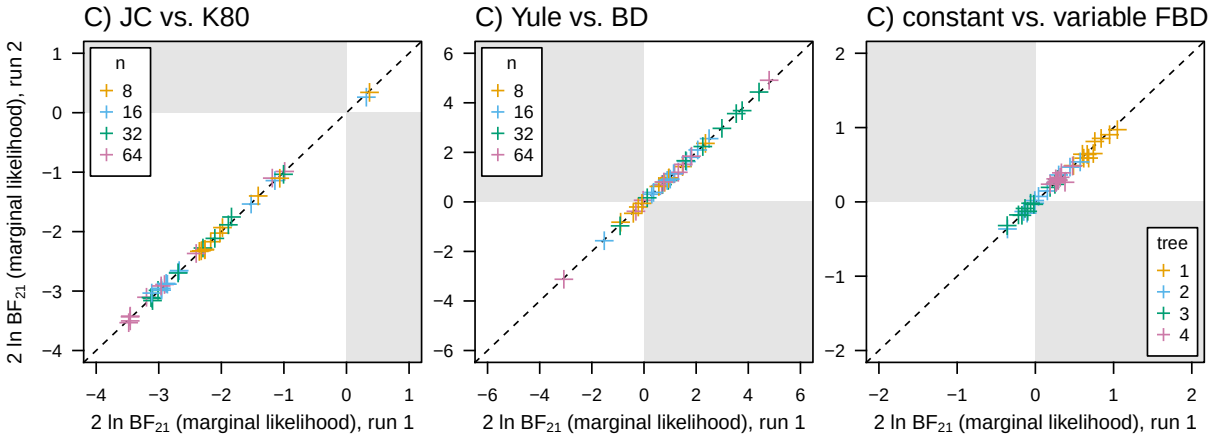

Figure S1: Precision of Bayes factor calculation using marginal likelihoods. Each analysis was performed twice, and the value from one run (x-axis) is plotted against the second run (y-axis). A) Bayes factors between JC and K80 models. B) Bayes factors between Yule and birth-death models. C) Bayes factors between a model with constant fossilization rates and one with variable fossilization rates. (see caption of Fig. 1)

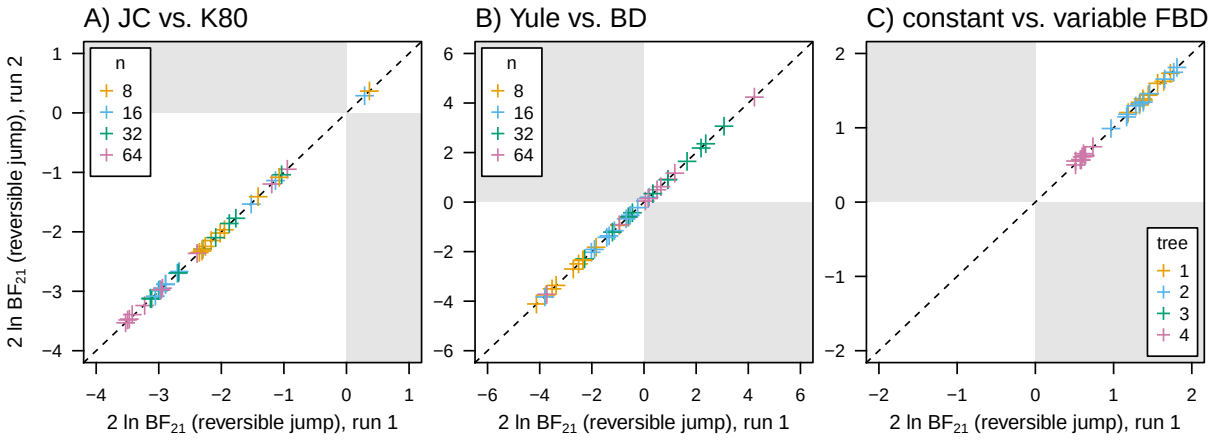

Figure S2: Precision of Bayes factor calculation using reversible-jump MCMC. Each analysis was performed twice, and the value from one run (x-axis) is plotted against the second run (y-axis). A) Bayes factors between JC and K80 models. B) Bayes factors between Yule and birth-death models. C) Bayes factors between a model with constant fossilization rates and one with variable fossilization rates. (see caption of Fig. 1)

S§2 Empirical Analysis

S§2.1 Analysis

We re-analyzed the empirical dataset of marattialean ferns from our previous study (May et al. 2021) to demonstrate the impact of not treating samples as data in a realistic model-comparison scenario, as well as to provide an example of the impact of the tree model on estimates of divergence times. For the sake of computational tractability, we included only ingroup taxa—comprising 26 extant and 45 extinct samples—and only analyzed the binary morphological data from that study (*i.e.*, we excluded the molecular data and multistate morphological characters).

We analyzed this dataset under an Mkv model (Lewis 2001) with gamma-distributed rate variation among characters (Yang 1994), and assumed that rates of morphological evolution varied across branches of the tree according to an uncorrelated-lognormal relaxed-clock model (UCLN Drummond et al. 2006). We compared the fit of two competing birth-death models: 1) a fossilized birth-death model where speciation and extinction rates varied over time, but the fossilization rate was constant over time (the "constant" model), and; 2) a fossilized birth-death model where fossilization, speciation, and extinction rates each varied over time (the "variable" model). Specifying arbitrary variable-rate fossilized birth-death models that are amenable to efficient reversible-jump MCMC is non-trivial. We therefore used the results from our previous study to constrain the rate variation so that it was both appropriate for the dataset, and possible to specify in the existing reversible-jump machinery available in `RevBayes`. For the speciation- and extinction-rate variation, we assumed that these rates varied according to a piecewise-constant model defined by five time intervals intended to capture the major patterns present in our prior analyses: $(\infty, 323.2]$, $(323.2, 298.9]$, $(298.9, 66.0]$, $(66.0, 5.3]$, $(5.3, 0.0]$. Within each time interval, speciation and extinction rates were drawn independently from a shared prior distribution. For the model that allowed the fossilization rate to vary, we assumed a piecewise constant model with four time intervals: $(\infty, 323.2]$, $(323.2, 298.9]$, $(298.9, 66.0]$, $(66.0, 0.0]$, with corresponding fossilization rates $\{\psi_1, \psi_2, \psi_3, \psi_4\}$. We assumed that the rates of the first and fourth interval were the same ($\psi_1 = \psi_4$), but allowed the fossilization rates for the second and third intervals to be different, reflecting an apparent peak in fossilization rates in the Pennsylvanian (the second interval), followed by moderate fossilization rates from the Permian to the end of the Cretaceous (the third interval).

For both tree models, we fixed the tree topology to the maximum-clade-credibility (MCC) tree topology inferred in our previous study, but estimated the node ages, the fossilized birth-death parameters, the character-evolution parameters, and also whether each fossil was a sampled ancestor. We then computed the Bayes factors between the constant and variable models using SS and RJ MCMC. Our empirical data and code for these analyses (including specific parameter and prior settings) are available at `XXXX`.

S§2.2 Results

Under the variable model, we infer extreme variation in fossilization rates: rates are inferred to be substantially higher during the Pennsylvanian than in the other time intervals, and rates from the Permian to the Late Cretaceous are also elevated compared to the first and fourth intervals (Fig. S3). This result is unsurprising, given that 24 of the 45 fossil samples come from the 23 My window that constitutes the Pennsylvanian subperiod. Despite this evident rate variation, BFs based on marginal likelihoods favor the constant model (2 ln BF $\approx$ 3, Tables S.1 and S.2). By contrast, BFs based on posterior model probabilities decisively favor the variable model (2 ln BF $\approx$ 18, Table S.2). In other words, conventional marginal-likelihood-based BFs incorrectly indicate strong evidence for a decisively worse model.

Table S.1: Marginal likelihoods for the constant model computed with stepping-stone sampling. We performed four independent runs to ensure precise marginal likelihood estimates (runs 1 through 4); we report the mean and standard error of the mean (final column).

Model	Run 1	Run 2	Run 3	Run 4	Mean ($\pm$ SD)
Constant	-883.0418	-882.8819	-882.9806	-882.9524	-882.9602 (± 0.03634)
Variable	-884.5965	-884.5084	-884.4403	-884.4645	-884.5025 (± 0.03136)

Table S.2: 2 ln Bayes factors between constant and variable models, computed using two methods. We performed four independent runs to ensure precise Bayes factor estimates (runs 1 through 4); we report the mean and standard error of the mean (final column).

Method	Run 1	Run 2	Run 3	Run 4	Mean ($\pm$ SD)
SS	-3.0770	-3.3178	-2.9194	-3.0242	-3.0846 (± 0.0843)
RJ	18.1055	17.9870	18.0041	18.0804	18.0442 (± 0.0287)

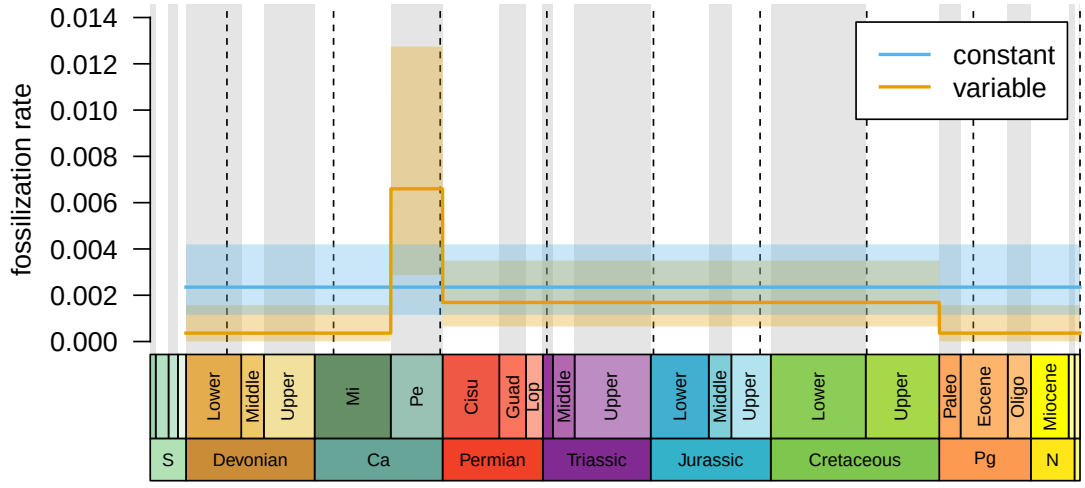

Figure S3: Estimated fossilization rates under the constant fossilization-rate model (blue) and the variable fossilization-rate model (orange). Dark lines correspond to the mean posterior rate at each time point, and colored regions correspond to the 95% credible interval. Dashed lines are placed at 50 My intervals.

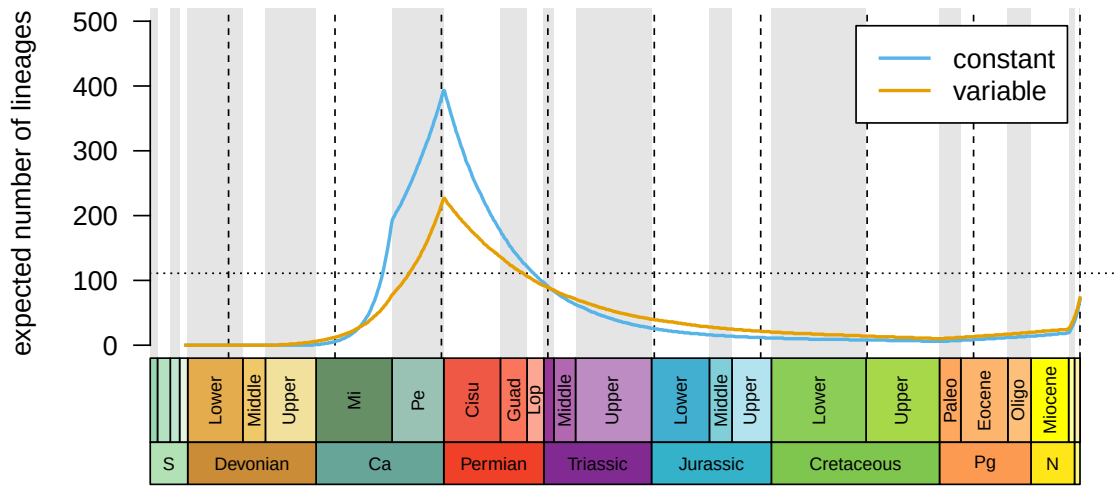

Figure S4: Median number of lineages over time, as predicted by the constant fossilization-rate model (blue) and the variable fossilization-rate model (orange). Vertical dashed lines are placed at 50 My intervals; horizontal dashed line is the number of extant lineages.

The different models produce significantly different inferences about the history of diversity for the group. We simulated 10,000 histories of lineage diversification under each model, and discretized time into many (10,000) small time intervals. We then computed the median number of lineages alive in each time interval over the group's history (Fig. S4). Under M_1, we predict an average peak diversity of ≈ 945 lineages in the Pennsylvanian; by contrast, M_2 predicts ≈ 546 lineages at that time. This discrepancy likely reflects the fact that M_1 requires a larger number of lineages during the Pennsylvanian in order to preserve the observed number of samples, given that fossilization rates are lower at that time relative to M_2 (Fig. S3).

Beside providing qualitatively different inferences about the nature of the fossilization process and the underlying history of diversification, the two tree models strongly influence divergence-time estimates for this dataset. We computed the mean of the posterior distribution of the age of each node in the tree under both models as $\Delta = |a_v - a_c|$, where a_c and a_v represent the posterior-mean estimate under the constant and variable models, respectively. Posterior-mean estimates of divergence times under the constant model differ substantially from those under the variable model: some nodes ages are different by 15 million years between the two models (Fig. S5, left). Divergence-time estimates for young nodes are systematically more recent under the constant model, *i.e.*,, the younger nodes are disproportionately pulled toward the present (Fig. S6). Presumably, this pattern reflects the fact that the fossilization rates in the Cenozoic are higher under the constant model than the variable model (Fig. S3): when fossilization rates are high, older clades in these time intervals imply more missing fossils, and are therefore "penalized" by the model (*i.e.*, younger clades are preferred).

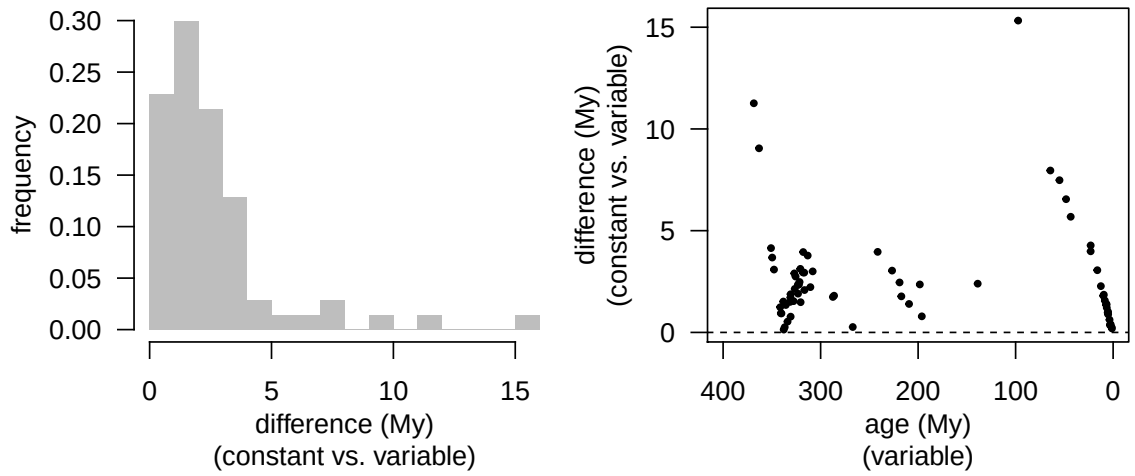

Figure S5: Left) Absolute differences in posterior mean node-age estimates under the two tree models. Right) Absolute differences in posterior mean node-age estimates under the two tree models as a function of age (the mean age estimated under the variable model).

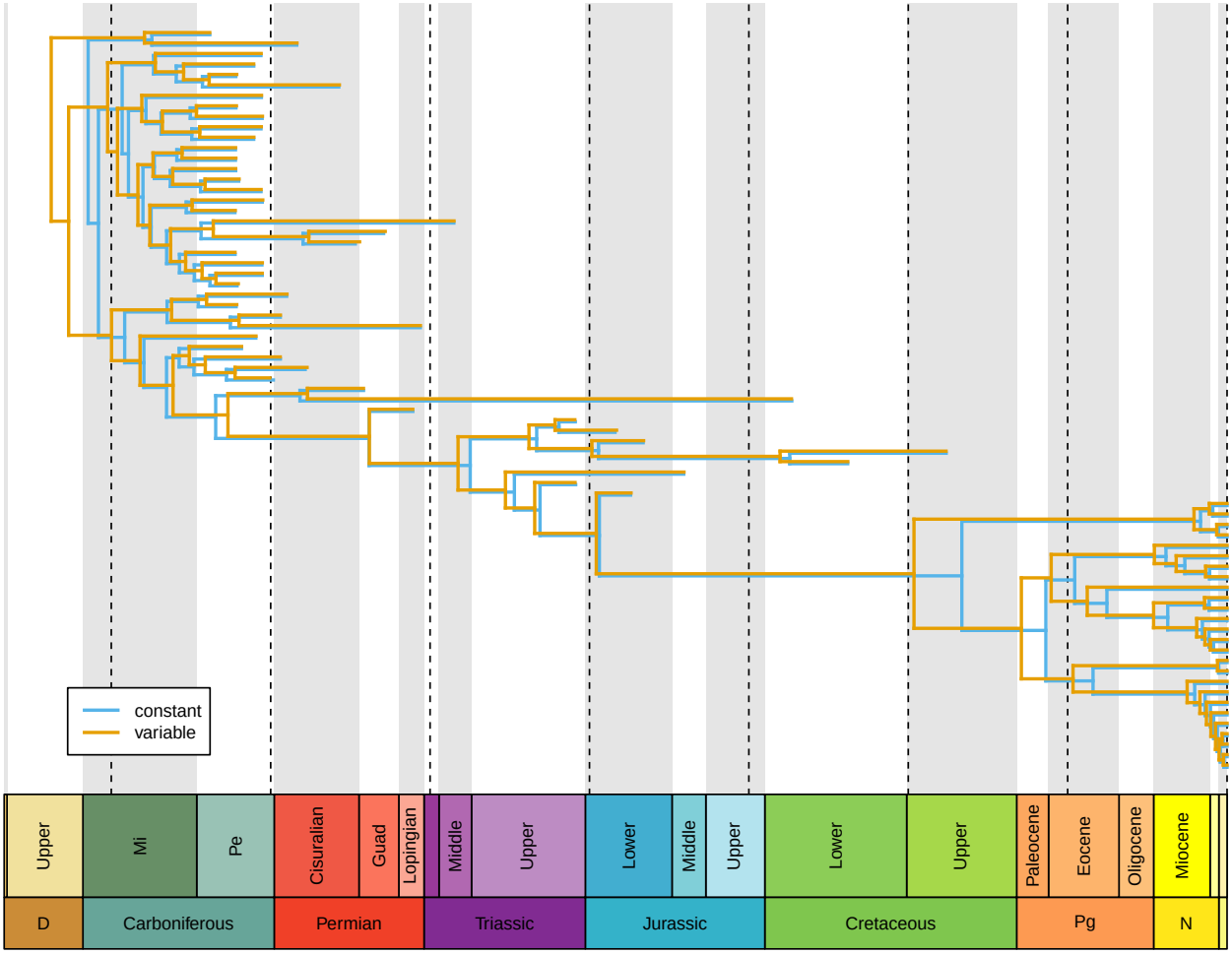

Figure S6: Phylogenies estimated under the constant (blue) and variable (orange) fossilized birth-death model. Node ages correspond to the the posterior mean age inferred under the model. Dashed lines are placed at 50 My intervals.

S§3 Samples As Data

S§3.1 The standard phylogenetic model

In a standard fossilized birth-death analysis (Heath et al. 2014; Zhang et al. 2016), we imagine that a phylogeny, Ψ, evolves under a fossilized birth-death process with a set of parameters θ_Ψ, which consists of speciation, extinction, and fossilization rates (λ, μ, and ϕ, respectively), and a sampling probability for extant species, ρ. Additionally, we imagine that a set of characters (typically a combination of molecular and discrete morphological characters) evolve along the branches of the tree according to a defined model with a set of parameters θ_x. These processes give rise to a character dataset, X.

Given a particular character dataset, X, our goal is to infer the phylogeny and model parameters that gave rise to that dataset. To do so, we can apply Bayes' theorem:

$$P(\Psi, \theta_x, \theta_\Psi \mid X) = \frac{\overbrace{P(X \mid \Psi, \theta_x)}^{\text{likelihood}} \overbrace{P(\Psi \mid \theta_\Psi) P(\theta_x) P(\theta_\Psi)}^{\text{priors}}}{\underbrace{P(X)}_{\text{marginal likelihood}}}. \tag{S.1}$$

The prior distributions $P(\theta_x)$ and $P(\theta_\Psi)$ represent our belief about these model components before observing the data. $P(\Psi \mid \theta_\Psi)$ is generally considered a prior distribution on the phylogeny, and is the probability density of a "ranked, labeled[1]" tree (see equation [2] in Gavryushkina et al. 2014). These prior distributions are updated by information in the character data by the likelihood function, $P(X \mid \Psi, \theta_x)$, to produce our posterior belief about the tree and model parameters, $P(\Psi, \theta_x, \theta_\Psi \mid X)$. The denominator—$P(X)$, *i.e.*, the marginal likelihood—is the likelihood function averaged over all of the model parameters in proportion to their prior probability:

$$P(X) = \iiint P(X \mid \Psi, \theta_x) P(\Psi \mid \theta_\Psi) P(\theta_x) P(\theta_\Psi) \, d\theta_x \, d\theta_\Psi \, d\Psi \tag{S.2}$$

(The integrals here represent multidimensional integration over θ_x and θ_Ψ, as well as summation over all possible tree topologies and integration over all sets of branch lengths in Ψ; we omit the domains of integration throughout this document for the sake of simplicity.) Different models will have different marginal likelihoods; we indicate the marginal likelihood for model i as $P_i(X)$ when we need to distinguish among models.

Typically, we refer to probabilities of data (observations) as likelihoods (technically, in a frequentist framework, the likelihood of the parameters is proportional to the probability of the data given the parameters), and probabilities of parameters as prior probabilities. The labeling of terms affects marginal likelihoods estimated using standard methods, which depend on treating likelihoods and priors differently (*e.g.*, Lartillot and Philippe 2006; Xie et al. 2011). However, the number of samples and their ages are observations that provide information about the lineage-diversification process, independent from the character data; indeed, paleontologists regularly use this type of information, by itself, to estimate parameters of lineage-diversification models. From this perspective, the "prior probability" of the phylogeny, $P(\Psi \mid \theta_\Psi)$, actually represents the joint probability of the samples and the phylogeny, and some of this probability—the portion related to the samples—belongs to the likelihood.

[1] Here, "ranked" means the nodes in the tree have ages, and "labeled" means the tips and sampled ancestors in the tree are associated with specific named samples in our dataset (Murtagh 1984).

S§3.2 A paleontological model

There is a long history of studying lineage diversification from a purely paleontological perspective (*e.g.*, Raup 1975, 1985; Foote 2000, 2001). Some of these methods, particularly `PyRates` (Silvestro et al. 2014, 2019), use stochastic birth-death models that are effectively interchangeable with those used in phylogenetic methods. In this framework, the data are taken to be fossil occurrences of the clade(s) of interest, and the goal is to estimate speciation (origination) and extinction rates based on how the fossil occurrences are distributed over time. In contrast to phylogenetic methods, the relationships among the fossil occurrences are not of direct interest, so this method does not depend on character data or models of character evolution.

In a Bayesian framework, we can conceive of a generic "paleontological" model (similar to `PyRates`) that would be represented as:

$$P(\lambda, \mu, \phi, \rho \mid S) = \frac{\overbrace{P(S \mid \lambda, \mu, \phi, \rho)}^{\text{likelihood}} \overbrace{P(\lambda)P(\mu)P(\phi)P(\rho)}^{\text{priors}}}{\underbrace{P(S)}_{\text{marginal likelihood}}}, \tag{S.3}$$

where S represents the set of samples, comprising individual fossil occurrences—the ages of the specimens, which for the time being we assume are known exactly, *i.e.*, there is no stratigraphic-age uncertainty—as well as any extant members of the group being analyzed, and the diversification parameters are the same as those described above for the fossilized birth-death model. Here, it is clear that $P(S \mid \lambda, \mu, \phi, \rho)$ is the probability of the observations, and is thus (proportional to) the likelihood function. It may be difficult to compute the probability of the samples without knowing the complete tree (including unsampled lineages), $\widetilde{\Psi}$. However, if we can compute the conditional probability of the samples given the complete tree, in principle we can write the unconditional probability of the samples as:

$$\begin{aligned} \overbrace{P(S \mid \lambda, \mu, \phi, \rho)}^{\substack{\text{probability of} \\ \text{the samples}}} &= \int \overbrace{P(S, \widetilde{\Psi} \mid \lambda, \mu, \phi, \rho)}^{\substack{\text{joint probability of} \\ \text{the tree and samples}}} d\widetilde{\Psi} \\ &= \int \underbrace{P(S \mid \widetilde{\Psi}, \phi, \rho)}_{\substack{\text{conditional probability} \\ \text{of the samples,} \\ \text{given the tree}}} \underbrace{P(\widetilde{\Psi} \mid \lambda, \mu)}_{\substack{\text{probability of} \\ \text{the tree}}} d\widetilde{\Psi}, \end{aligned} \tag{S.4}$$

where the integral represents integration over all possible complete phylogenies, $\widetilde{\Psi}$, in proportion to their probability. Given the phylogeny, the probability of the samples is independent of the diversification parameters, λ and μ; likewise, the probability of the full tree does not depend on the sampling parameters, ϕ and ρ. (Note that this complete tree is different from the "reconstructed" tree in equation [S.1], Ψ, which hides all the unsampled lineages.)

While the probability of the samples given the complete tree—$P(S \mid \widetilde{\Psi}, \phi, \rho)$—may be relatively easy to compute, the above integral may be quite difficult to compute. However, a simple solution would be to include the complete phylogeny in the posterior distribution as an additional parameter:

$$P(\widetilde{\Psi}, \lambda, \mu, \phi, \rho \mid S) = \frac{P(S \mid \widetilde{\Psi}, \phi, \rho)P(\widetilde{\Psi} \mid \lambda, \mu)P(\lambda)P(\mu)P(\phi)P(\rho)}{P(S)}. \tag{S.5}$$

Importantly, this will produce the same posterior estimates of λ, μ, ϕ, and ρ as produced by equation (S.3) when we integrate over the phylogeny:

$$\int P(\widetilde{\Psi}, \lambda, \mu, \phi, \rho \mid S)\, d\widetilde{\Psi} = \int \frac{P(S \mid \widetilde{\Psi}, \phi, \rho) P(\widetilde{\Psi} \mid \lambda, \mu) P(\lambda) P(\mu) P(\phi) P(\rho)}{P(S)}\, d\widetilde{\Psi}$$

$$P(\lambda, \mu, \phi, \rho \mid S) = \frac{\left[\int P(S \mid \widetilde{\Psi}, \phi, \rho) P(\widetilde{\Psi} \mid \lambda, \mu)\, d\widetilde{\Psi}\right] P(\lambda) P(\mu) P(\phi) P(\rho)}{P(S)}$$

$$P(\lambda, \mu, \phi, \rho \mid S) = \frac{P(S \mid \lambda, \mu, \phi, \rho) P(\lambda) P(\mu) P(\phi) P(\rho)}{P(S)}.$$

Additionally, both representations have the same marginal likelihood, $P(S)$. The equation (S.5) implies an approach that is very similar to the approach used by the Bayesian program `PyRates` (Silvestro et al. 2014), which estimates the times of origin and extinction of every sampled lineages (much like branch lengths in the complete phylogeny), but not the relationships between lineages (*i.e.*, the phylogenetic topology).

S§3.3 Combining phylogenetic and paleontological models

While the phylogenetic and paleontological models use the same underlying models of lineage diversification and sampling, they do not appear to use the same data (the phylogenetic model uses the character data, X, whereas the paleontological model uses the sample data, S). We can resolve this apparent discrepancy by simply adding character data, X, and a model of character evolution to the paleontological model that includes an implicit tree (equation [S.5]). In this case, the data are both S and X, and the corresponding posterior distribution is:

$$P(\widetilde{\Psi}, \theta_x, \lambda, \mu, \phi, \rho \mid S, X) = \frac{\overbrace{P(X \mid \widetilde{\Psi}, \theta_x) P(S \mid \widetilde{\Psi}, \phi, \rho)}^{\text{likelihood}} \overbrace{P(\widetilde{\Psi} \mid \lambda, \mu) P(\theta_x) P(\lambda) P(\mu) P(\phi) P(\rho)}^{\text{priors}}}{\underbrace{P(S, X)}_{\text{marginal likelihood}}}.$$

We can attempt to derive something like the standard phylogenetic model—equation (S.1)—from this combined model by first recognizing $P(S \mid \widetilde{\Psi}, \phi, \rho) P(\widetilde{\Psi} \mid \lambda, \mu)$ as the joint probability of the samples and full tree, $P(S, \widetilde{\Psi} \mid \lambda, \mu, \phi, \rho)$ (see equation [S.4]):

$$P(\widetilde{\Psi}, \theta_x, \lambda, \mu, \phi, \rho \mid S, X) = \frac{\overbrace{P(X \mid \widetilde{\Psi}, \theta_x)}^{\text{likelihood}} \overbrace{P(S, \widetilde{\Psi} \mid \lambda, \mu, \phi, \rho)}^{\text{some likelihood, some prior}} \overbrace{P(\theta_x) P(\lambda) P(\mu) P(\phi) P(\rho)}^{\text{prior}}}{P(S, X)}.$$

Next, we must reduce $\widetilde{\Psi}$ to Ψ. For any given Ψ, there are an infinite number of unobserved histories consistent with Ψ, each of which produces a unique $\widetilde{\Psi}$. We label the unobserved history Ψ^c, and say that $\widetilde{\Psi} = \{\Psi, \Psi^c\}$. The posterior distribution of the phylogenetic model should integrate over all

possible unobserved histories in proportion to their probability:

$$
\begin{aligned}
P(\Psi, \theta_x, \lambda, \mu, \phi, \rho \mid S, X) &= \int P(\Psi, \Psi^c, \theta_x, \lambda, \mu, \phi, \rho \mid S, X)\, d\Psi^c \\
&= \frac{\left[\int P(X \mid S, \Psi, \Psi^c, \theta_x) P(S, \Psi, \Psi^c \mid \lambda, \mu, \phi, \rho)\, d\Psi^c\right] P(\theta_x)P(\lambda)P(\mu)P(\phi)P(\rho)}{P(S, X)} \\
&= \frac{P(X \mid S, \Psi, \theta_x) P(S, \Psi \mid \lambda, \mu, \phi, \rho) P(\theta_x)P(\lambda)P(\mu)P(\phi)P(\rho)}{P(S, X)} \\
P(\Psi, \theta_x, \theta_\Psi \mid S, X) &= \frac{\overbrace{P(X \mid S, \Psi, \theta_x)}^{\text{likelihood}} \overbrace{P(S, \Psi \mid \theta_\Psi)}^{\substack{\text{some likelihood,} \\ \text{some prior}}} \overbrace{P(\theta_x)P(\theta_\Psi)}^{\text{prior}}}{\underbrace{P(S, X)}_{\text{marginal likelihood}}}. \qquad \text{(S.6)}
\end{aligned}
$$

(We include S as a dependency in $P(X \mid S, \Psi, \theta_x)$ because Ψ is a function of S and $\widetilde{\Psi}$.)

Equation (S.6) is very similar to the standard phylogenetic representation, but with critical differences. First, the data include both X and S, rather than just X; consequently, the marginal likelihoods must be different. Second, the prior probability of the sampled tree—$P(\Psi \mid \theta_\Psi)$ in equation (S.1)—has been replaced with the joint probability of the samples and the sampled tree, $P(S, \Psi \mid \theta_\Psi)$. Because this joint probability includes observations, some part of it should be regarded as part of the likelihood of the model.

These equations demonstrate that probabilities we are used to thinking of as prior probabilities—specifically, the probabilities of trees under a birth-death model—are actually an ambiguous mixture of likelihood-like and prior-like quantities. That is, the likelihood and prior functions in the standard Bayesian model are mislabeled. We explore the consequences of this mislabeling in Section S§4.

S§3.4 When sample ages are uncertain

So far, we have assumed that the ages of the fossil occurrences are known without error, which helps to clarify our main argument that samples should be treated as data. However, in real datasets, the ages of fossil specimens are often uncertain, because the age of the sediments in which the specimens are found can only be known within a certain interval. This phenomenon—referred to as stratigraphic-age uncertainty—is somewhat orthogonal to our argument, but we mention it here because previous work has argued that stratigraphic-age uncertainty should be treated a part of the likelihood function Drummond and Stadler (2016). We agree with this perspective, and show how it fits in to the framework we outlined above. Unfortunately, stratigraphic-age uncertainty leads to additional challenges when computing marginal likelihoods.

Following Drummond and Stadler (2016), we represent stratigraphic-age data as $A = \{a_1, a_2, \ldots, a_n\}$ for the n samples, where $a_i = \{\overset{\vee}{a}_i, \hat{a}_i\}$ are the minimum and maximum ages of the i^{th} sample, respectively. The probability of the data (the stratigraphic ranges) would then be:

$$
\begin{aligned}
P(A \mid \lambda, \mu, \phi, \rho) &= \int P(A, S \mid \lambda, \mu, \phi, \rho)\, dS \\
&= \int P(A \mid S) P(S \mid \lambda, \mu, \phi, \rho)\, dS, \qquad \text{(S.7)}
\end{aligned}
$$

where the integration is over the exact ages of all the samples, S, and $P(A \mid S)$ is a product of indicator

functions:

$$P(A \mid S) = \prod_{i}^{n} P(a_i \mid S_i),$$

with

$$P(a_i \mid S_i) = \begin{cases} 1 & \text{if } \overset{\vee}{a}_i \leq s_i \leq \hat{a}_i \\ 0 & \text{otherwise.} \end{cases}$$

Including the full tree, the likelihood with stratigraphic age uncertainty becomes:

$$P(A \mid \lambda, \mu, \phi, \rho) = \iint P(A \mid S) P(S \mid \widetilde{\Psi}, \phi, \rho) P(\widetilde{\Psi} \mid \lambda, \mu)\, d\widetilde{\Psi}\, dS.$$

Even if this integral were analytically tractable, we could not use it with character data, because the probability of the character data will generally depend on the exact ages of the samples. However, we can use data augmentation (Tanner and Wong 1987) to include the exact ages in the model, and write the full posterior distribution:

$$P(S, \Psi, \theta_x, \theta_\Psi \mid A, X) = \frac{P(X \mid S, \Psi, \theta_x) P(A \mid S) P(S, \Psi \mid \theta_\Psi) P(\theta_x) P(\theta_\Psi)}{P(A, X)},$$

where now the data are A and X. This approach amounts to a data augmentation because the likelihood should average over the exact sample ages, S, as implied by equation (S.7).

Without special machinery, generic methods for computing the marginal likelihood that depend on raising the likelihood function to a power cannot effectively deal with data-augmented models (Rodrigue and Aris-Brosou 2011). As a consequence, correct solutions for marginal-likelihood estimators with stratigraphic uncertainty are currently unavailable.

S§4 Sequential Bayesian Inference

Above, we showed that the traditional phylogenetic model mistreats the probability of the samples as part of the prior rather than the likelihood function. Here, we use the principle of of sequential Bayesian inference to understand the quantitative consequences of this error. The parameters of the prior distributions (hyperparameters) we choose for a Bayesian model represent our prior belief about plausible parameter values, and in principle reflect our previous experiences with analyzing relevant data (or ignorance, if we have no previous experience). In a sense, when informed by previous analysis, these hyperparameters encapsulate the information in the previous datasets about the parameters, *i.e.*, they can be viewed as "old" data. When we analyze a "new" dataset, we update our prior beliefs accordingly. We can repeat this process indefinitely, as we collect additional datasets. This sequential Bayesian updating process is the basis of Lindley's aphorism that "today's posterior is tomorrow's prior" (Lindley 1972).

When we perform a Bayesian phylogenetic analysis under a birth-death model, we can imagine collecting two datasets. We first collect samples, represented by their ages S. We may infer the tree model (birth-death and sampling) parameters, θ_Ψ, directly from this dataset. We can write the posterior distribution of this model as:

$$P(\theta_\Psi \mid S) = \frac{\overbrace{P(S \mid \theta_\Psi)}^{\text{likelihood}}\overbrace{P(\theta_\Psi)}^{\text{prior}}}{\underbrace{P(S)}_{\substack{\text{marginal likelihood of} \\ \text{the samples}}}}, \tag{S.8}$$

which corresponds to the posterior distribution of a "paleontological" model (equation [S.3]).

However, if we then become additionally interested in the phylogenetic relationships themselves, we can assemble a character dataset, X. Rather than re-doing the initial analysis, we may apply the principle of sequential Bayesian updating and use the first posterior as a prior in our second analysis:

$$P(\Psi, \theta_x, \theta_\Psi \mid X, S) = \frac{\overbrace{P(X \mid S, \Psi, \theta_x)}^{\text{likelihood}}\overbrace{P(\Psi \mid \theta_\Psi)P(\theta_\Psi \mid S)P(\theta_x)}^{\text{priors}}}{\underbrace{P(X \mid S)}_{\substack{\text{marginal likelihood of the characters} \\ \text{given the samples}}}}. \tag{S.9}$$

In this equation, S is effectively treated as a hyperparameter, *i.e.*, a fixed parameter of the prior distribution on θ_Ψ. (We note that *all* prior distributions have hyperparameters, but we usually exclude them from our notation for simplicity.)

Alternatively, we could start again and do both analyses simultaneously (jointly). The posterior of such a joint analysis would be

$$P(\Psi, \theta_x, \theta_\Psi \mid X, S) = \frac{\overbrace{P(X \mid S, \Psi, \theta_x)}^{\text{likelihood}}\overbrace{P(S, \Psi \mid \theta_\Psi)}^{\substack{\text{some likelihood,} \\ \text{some prior}}}\overbrace{P(\theta_\Psi)P(\theta_x)}^{\text{priors}}}{\underbrace{P(X, S)}_{\substack{\text{marginal likelihood of} \\ \text{samples and characters}}}}, \tag{S.10}$$

where we view the likelihood function as $P(X \mid \Psi, \theta_x)$, as well as some contribution from $P(S, \Psi \mid \theta_\Psi)$, as we explain in Section S§3. We can verify that the posterior distribution from the joint analysis, equation (S.10), is equivalent to the posterior distribution after the second step of the sequential analysis

by substituting equation (S.8) into equation (S.9), and recognizing that

$$P(X, S) = P(X \mid S)P(S),$$

i.e., that the marginal likelihood of the joint analysis is the product of the marginal likelihoods of each step in the sequential analysis.

The remaining task is to explain why SS and RJ MCMC estimate different Bayes factors. Methods for calculating the marginal likelihood, such as SS, require that we clearly distinguish the probability terms that are "likelihood" from those that are "prior". In the standard phylogenetic notation (equation [S.1]), the probability of the character data is labeled the likelihood, while the joint probability of the tree and samples is labeled the prior. This corresponds to the labeling in the second step of the sequential analysis, equation (S.9), in which case the marginal likelihood is $P(X \mid S)$; this marginal likelihood perceives the samples as "old" data, and only computes the marginal likelihood of the "new" data, X. When we compare two models in this way, we are essentially imagining that for each model, we first update the priors according to S, and then compute the marginal likelihood of X given the corresponding posteriors from the first step. The resulting Bayes factor between the two models i and j will be

$$\mathrm{BF}_{ij} = \frac{P_i(X \mid S)}{P_j(X \mid S)}, \tag{S.11}$$

where $P_k(X \mid S)$ is the marginal likelihood of model k. By contrast, RJ MCMC does not depend on the labeling of probability terms. In this case, the dataset implicitly includes both samples and character data, and the Bayes factors will be

$$\mathrm{BF}_{ij} = \frac{P_i(X, S)}{P_j(X, S)} = \frac{P_i(X \mid S)}{P_j(X \mid S)} \frac{P_i(S)}{P_j(S)}, \tag{S.12}$$

Equations (S.11) and (S.12) predict that the discrepancy in Bayes factors between SS and RJ MCMC should be equal to the ratio of the marginal likelihoods of the samples, $P_i(S) \div P_j(S)$.

For birth-death processes that generate contemporaneous samples, the probability of the samples for a given set of parameters is straightforward to compute (Kendall 1948; Höhna 2015), and we can relatively easily compute the marginal likelihood. Indeed, when we calculate the marginal likelihood of the samples, it corresponds exactly to the discrepancy we observe between SS and RJ MCMC estimates of the Bayes factor (Fig. 1B, main text, middle, dashed lines). For birth-death processes generating non-contemporaneous samples, the marginal probability of the samples is not possible to compute analytically, and therefore we cannot make precise numerical predictions about the discrepancy.

S§5 Factorizing Bayes' Theorem

The problem with equation (S.6) is that $P(S, \Psi \mid \theta_\Psi)$ combines likelihood and prior quantities. This could be resolved by factoring this quantity as:

$$P(S, \Psi \mid \theta_\Psi) = P(\Psi \mid S, \theta_\Psi) P(S \mid \theta_\Psi),$$

in which case each term on the right is unambiguously likelihood (the marginal probability of the samples) or not (the conditional probability of the tree, given the samples). This can be directly substituted into the posterior:

$$P(\Psi, \theta_x, \theta_\Psi \mid S, X) = \frac{P(X \mid S, \Psi, \theta_x) P(\Psi \mid S, \theta_\Psi) P(S \mid \theta_\Psi) P(\theta_x) P(\theta_\Psi)}{P(S, X)}, \tag{S.13}$$

which is compatible with standard numerical methods for computing the marginal likelihood that rely on raising the likelihood function to a power ("power-posterior" methods, for example, path-sampling and stepping-stone-sampling algorithms; Lartillot and Philippe 2006; Xie et al. 2011). However, the marginal probability of the samples and the conditional probability of the tree and the samples are not generally easy to compute; in particular, analytical solutions are only currently available for simple models of contemporaneous samples, and may be impossible for more complex models.

We implemented this solution for Yule and birth-death models producing contemporaneous samples. For these models, $P(S \mid \theta_\Psi)$ is the probability of realizing n samples of a given age, for which there are available analytical solutions (*e.g.*, equation [8] from Höhna 2015). Likewise, $P(\Psi \mid S, \theta_\Psi)$—the probability of the tree conditional on n samples—also has an available analytical solution (*e.g.*, equation [3] from Yang and Rannala 1997). We re-analyzed our simulated data from Section S§1.2 using this formulation, demonstrating that it provides correct marginal-likelihood estimates (Fig. S7).

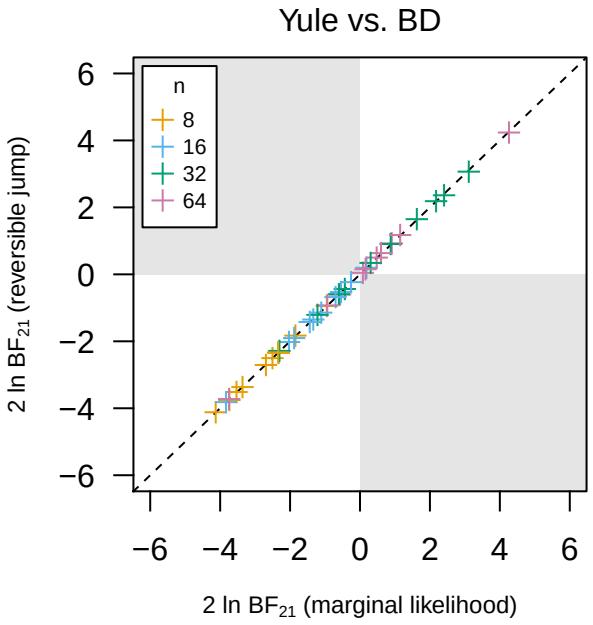

Figure S7: Bayes factor discrepancies are resolved by refactoring Bayes' theorem. We compared the fit of two birth-death processes—the Yule model (with no extinction rate parameter) and the standard birth-death (BD) model—to datasets simulated under the BD model (as described in Section S§1.2). We corrected the likelihood function according to equation (S.13). As expected, there is no discrepancy between the BFs calculated using marginal likelihoods and reversible-jump MCMC.

S§6 Posterior-Predictive Simulation with Samples as Data

While Bayes factors are useful for comparing the *relative* fit of competing models, they provide no guarantee that the best model adequately describes the process that gave rise to the observed data. Posterior-predictive simulation (PPS; Gelman et al. 1996) is a Bayesian tool that fills this gap by assessing model adequacy—whether our inference model provides an adequate description of the true process that produced our observed dataset—and is therefore useful for assessing *absolute* model fit. Generally, the procedure works by drawing parameters of the model from their joint posterior distribution (*e.g.*, as produced by an MCMC analysis), simulating new datasets under these parameters, and checking whether the simulated data resembles the observed dataset: are the values of a particular summary statistic computed from the simulated datasets reasonably close to the value of that statistic computed from the empirical data?

In phylogenetics, the PPS has been largely limited to morphological or molecular character datasets (*e.g.*, Brown 2014; Höhna et al. 2018; Slater and Pennell 2014; May et al. 2021). This limited application of PPS is understandable, given that the character datasets are the only component of the study that is considered to be data under the standard phylogenetic model. However, for studies that rely on the tree model, such as diversification-rate analyses or divergence-time estimation, a more natural summary statistic would be one that relates to characteristics of the sample, rather than the morphological or molecular data. For example, if we wanted to assess the adequacy of a diversification model, we might use the number of samples at a particular time as a test statistic. The availability of this application of sample-based test statistics is one of the primary benefits—to theoreticians and empiricists alike—of recognizing the samples themselves as data that inform the tree model.

To demonstrate the utility of posterior-predictive distributions for samples under birth-death models, we applied this technique to the Marattiales analyses described above (S§2). We simulated datasets by simulating trees under the sampled fossilized birth-death model parameters, and keeping track of the number of fossils recorded in each geological epoch. The posterior-predictive distributions of the number of samples shows that the model with constant fossilizaton rates does a poor job of predicting the observed number of fossils in the Mississippian, Pennsylvanian, and Cisuralian (Fig. S8). By contrast, the model with variable fossilization rates does a much better job at predicting the number of fossils in these (and subsequent) intervals (Fig. S8, right). This result is concordant with our relative measures of model fit (using Bayes factors), which we report in the main text, and demonstrates that the variable-rate model is not only better-fitting than its constant-rate counterpart, but moreover that it is an adequate representation of the process that generated our data. (We present these results as a proof-of-concept rather than as a method: developing appropriate posterior-predictive methods is a significant task that requires validation and evaluation of statistic properties.)

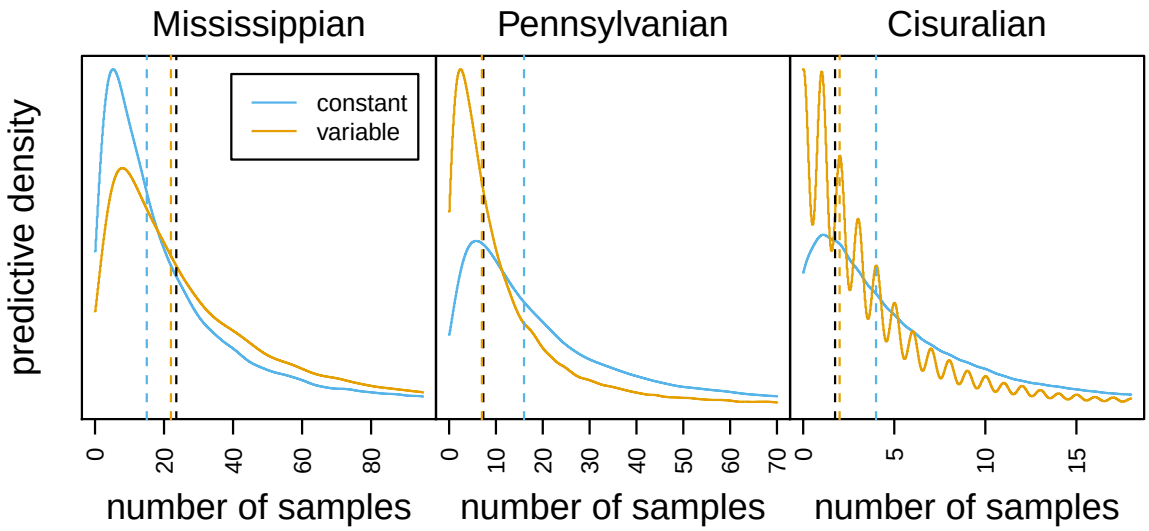

Figure S8: Posterior-predictive simulation for the Marattiales dataset under two models. We simulated fossil and extant marattialean datasets under a model with constant fossilization rates (blue) and variable fossilization rates (orange). Each density represents the posterior-predictive distribution for the number of samples in the given epoch. The black vertical line represent the observed number of samples in the epoch, and the colored vertical lines represent the median number of samples under the corresponding model.

S§7 Random Variables and Priors in Phylogenetic Inference

The problem that we detail in this manuscript—that the standard phylogenetic model mislabels samples as belonging to the prior rather than to the likelihood, which prevents accurate marginal likelihood calculation for tree models—is a specific example of a more general inconsistency in likelihood-based phylogenetics of distinguishing random variables (which have probabilities) from parameters (which have likelihoods). For example, there is a history of mistreating data as parameters in maximum-likelihood inference of ancestral states. Ancestral states are an outcome of the model: they are random variables, just as the character states at the tips are, an equivalence that is apparent when one recognizes that today's tip data are tomorrow's ancestral states. It is true, of course, that ancestral states are not observed, but that does not make them any less data-like; they are perhaps best conceived of as missing data, just as there can be missing data at the tips. There is nonetheless a strong history of treating ancestral states as parameters and, for example, comparing among different ancestral states with likelihood-ratio tests (*e.g.*, Pagel 1999). We contend that this approach is incorrect: ancestral states are not parameters and thus do not have likelihoods; rather, ancestral states are random variables to which we can assign different probabilities, given the tip data and model of character evolution (Yang et al. 1995; Yang 2014).

Similarly, there has been some historical disagreement about whether the phylogeny itself is a parameter or a random variable. The current dominant perspective, which derives from Felsenstein (Felsenstein 1973a,b), is that the tree is a parameter. In a maximum-likelihood framework, this perspective implies that the tree has a maximum-likelihood estimate; in a Bayesian framework, this perspective suggests that the tree should have a prior distribution. However, a minority perspective is that the tree should be viewed as a random variable just as the ancestral states are viewed as a random variable (Edwards 1970; Rannala and Yang 1996), and therefore should be associated with a probability distribution even in a maximum-likelihood framework. This latter perspective is even more germane today, considering the significant development of character-state-dependent diversification models (*i.e.*, the binary-state-specific speciation-and-extinction model [BiSSE; Maddison et al. 2007] and its derivatives). These models assume that rates of speciation and extinction are a function of an evolving character; since the evolution of the character is a random process, the resulting tree must be a random variable.

Our arguments are in alignment with the latter perspective: the tree should be viewed as the outcome of a random process, even in a maximum likelihood framework. However, while both Edwards (1970) and Rannala and Yang (1996) used birth-death models for the tree, they condition the model on achieving exactly the observed number of extant species. This is equivalent to performing the second step of a sequential Bayesian analysis, which we describe above. Our view is therefore an extension of Edwards', to include the samples themselves as part of the outcome.

References

Brown, J. M. (2014). Predictive approaches to assessing the fit of evolutionary models. *Systematic Biology*, 63(3):289–292.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):699–710.

Drummond, A. J. and Stadler, T. (2016). Bayesian phylogenetic estimation of fossil ages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1699):20150129.

Edwards, A. W. (1970). Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(2):155–164.

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

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

Foote, M. (2000). Origination and extinction components of taxonomic diversity: general problems. *Paleobiology*, 26(sp4):74–102.

Foote, M. (2001). Inferring temporal patterns of preservation, origination, and extinction from taxonomic survivorship analysis. *Paleobiology*, 27(4):602–630.

Friel, N., Hurn, M., and Wyse, J. (2014). Improving power posterior estimation of statistical evidence. *Statistics and Computing*, 24(5):709–723.

Gavryushkina, A., Welch, D., Stadler, T., and Drummond, A. J. (2014). Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Computational Biology*, 10(12):e1003919.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, pages 733–760.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Heath, T. A., Huelsenbeck, J. P., and Stadler, T. (2014). The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences, USA*, 111(29):E2957–E2966.

Höhna, S. (2015). The time-dependent reconstructed evolutionary process with a key-role for mass-extinction events. *Journal of Theoretical Biology*, 380:321–331.

Höhna, S., Coghill, L. M., Mount, G. G., Thomson, R. C., and Brown, J. M. (2018). P3: Phylogenetic posterior prediction in RevBayes. *Molecular Biology and Evolution*, 35(4):1028–1034.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65(4):726–736.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism*, 3:21–132.

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

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120.

Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6):913–925.

Lindley, D. V. (1972). *Bayesian statistics: A review*. SIAM.

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

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

Murtagh, F. (1984). Counting dendrograms: a survey. *Discrete Applied Mathematics*, 7(2):191–199.

Pagel, M. (1999). The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology*, 48(3):612–622.

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

Raup, D. M. (1975). Taxonomic survivorship curves and Van Valen's Law. *Paleobiology*, 1(1):82–96.

Raup, D. M. (1985). Mathematical models of cladogenesis. *Paleobiology*, 11(1):42–52.

Rodrigue, N. and Aris-Brosou, S. (2011). Fast Bayesian choice of phylogenetic models: Prospecting data augmentation–based thermodynamic integration. *Systematic Biology*, 60(6):881–887.

Silvestro, D., Salamin, N., Antonelli, A., and Meyer, X. (2019). Improved estimation of macroevolutionary rates from fossil data using a Bayesian framework. *Paleobiology*, 45(4):546–570.

Silvestro, D., Schnitzler, J., Liow, L. H., Antonelli, A., and Salamin, N. (2014). Bayesian estimation of speciation and extinction from incomplete fossil occurrence data. *Systematic Biology*, 63(3):349–367.

Slater, G. J. and Pennell, M. W. (2014). Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. *Systematic Biology*, 63(3):293–308.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, 60(2):150–160.

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.

Yang, Z. (2014). *Molecular Evolution: A Statistical Approach*. Oxford University Press.

Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141(4):1641–1650.

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

Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A., and Ronquist, F. (2016). Total-evidence dating under the fossilized birth–death process. *Systematic Biology*, 65(2):228–249.