Marginal likelihoods in phylogenetics: a review of methods and applications

Jamie R. Oaks*[1], Kerry A. Cobb[1], Vladimir N. Minin[2], and Adam D. Leaché[3]

[1]Department of Biological Sciences & Museum of Natural History, Auburn University, Auburn, Alabama 36849

[2]Department of Statistics, University of California, Irvine, California 92697

[3]Department of Biology & Burke Museum of Natural History and Culture, University of Washington, Seattle, Washington 98195

January 14, 2019

*Corresponding author: joaks@auburn.edu

Abstract

By providing a framework of accounting for the shared ancestry inherent to all life, phylogenetics is becoming the statistical foundation of biology. The importance of model choice continues to grow as phylogenetic models continue to increase in complexity to better capture micro and macroevolutionary processes. In a Bayesian framework, the marginal likelihood is how data update our prior beliefs about models, which gives us an intuitive measure of comparing model fit that is grounded in probability theory. Given the rapid increase in the number and complexity of phylogenetic models, methods for approximating marginal likelihoods are increasingly important. Here we try to provide an intuitive description of marginal likelihoods and why they are important in Bayesian model testing. We also categorize and review methods for estimating marginal likelihoods of phylogenetic models, highlighting several recent methods that provide well-behaved estimates. Furthermore, we review some empirical studies that demonstrate how marginal likelihoods can be used to learn about models of evolution from biological data. We discuss promising alternatives that can complement marginal likelihoods for Bayesian model choice, including posterior-predictive methods. Using simulations, we find one alternative method based on approximate-Bayesian computation (ABC) to be biased. We conclude by discussing the challenges of Bayesian model choice and future directions that promise to improve the approximation of marginal likelihoods and Bayesian phylogenetics as a whole.

KEY WORDS: phylogenetics, marginal likelihood, model choice

1 Introduction

Phylogenetics is rapidly progressing as the statistical foundation of comparative biology, providing a framework that accounts for the shared ancestry inherent in biological data. Soon after phylogenetics became feasible as a likelihood-based statistical endeavor (Felsenstein, 1981), models became richer to better capture processes of biological diversification and character change. This increasing trend in model complexity made Bayesian approaches appealing, because they can approximate posterior distributions of rich models by leveraging prior information and hierarchical models, where researchers can take into account uncertainty at all levels in the hierarchy.

From the earliest days of Bayesian phylogenetics (Rannala and Yang, 1996; Mau and Newton, 1997), the numerical tool of choice for approximating the posterior distribution was Markov chain Monte Carlo (MCMC). The popularity of MCMC was due, in no small part, to avoiding the calculation of the marginal likelihood of the model—the probability of the data under the model, averaged, with respect to the prior, over the whole parameter space. This marginalized measure of model fit is not easy to compute due to the large number of parameters in phylogenetic models (including the tree itself) over which the likelihood needs to be summed or integrated.

Nonetheless, marginal likelihoods are central to model comparison in a Bayesian framework. Learning about evolutionary patterns and processes via Bayesian comparison of phylogenetic models requires the calculation of marginal likelihoods. As the diversity and richness of phylogenetic models has increased, there has been a renewed appreciation of the importance of such Bayesian model comparison. As a result, there has been substantial work over the last decade to develop methods for estimating marginal likelihoods of phylogenetic models.

The goals of this review are to (1) try to provide some intuition about what marginal likelihoods are and why they can be useful, (2) review the various methods available for approximating marginal likelihoods of phylogenetic models, (3) review some of the ways

marginal likelihoods have been applied to learn about evolutionary history and processes, (4) highlight some alternatives to marginal likelihoods for Bayesian model comparison, (5) discuss some of the challenges of Bayesian model choice, and (6) highlight some promising avenues for advancing the field of Bayesian phylogenetics.

2 What are marginal likelihoods and why are they useful?

A marginal likelihood is the average fit of a model to a dataset. More specifically, it is an average over the entire parameter space of the likelihood weighted by the prior. For a phylogenetic model M with parameters that include the discrete topology (T) and continuous branch lengths and other parameters that govern the evolution of the characters along the tree (together represented by θ), the marginal likelihood can be represented as

$$p(D \mid M) = \sum_{T} \int_{\theta} p(D \mid T, \theta, M) p(T, \theta \mid M) \mathrm{d}\theta, \tag{1}$$

where D are the data. Each parameter of the model adds a dimension to the model, over which the likelihood must be averaged. The marginal likelihood is also the normalizing constant in the denominator of Bayes' rule that ensures the posterior is a proper probability density that sums and integrates to one:

$$p(T, \theta \mid D, M) = \frac{p(D \mid T, \theta, M) p(T, \theta \mid M)}{p(D \mid M)}. \tag{2}$$

Marginal likelihoods are the currency of model comparison in a Bayesian framework. This differs from the frequentist approach to model choice, which is based on comparing the maximum probability or density of the data under two models either using a likelihood ratio test or some information-theoretic criterion. Because adding a parameter (dimension) to a

model will always ensure a maximum likelihood at least as large as without the parameter, some penalty must be imposed when parameters are added. How large this penalty should be is not easy to define, which has led to many different possible criteria, e.g., the Akaike information criterion (AIC; Akaike, 1974), second-order AIC (AIC_C; Hurvich and Tsai, 1989; Sugiura, 1978), and Bayesian information criterion (BIC Schwarz, 1978).

Instead of focusing on the maximum likelihood of a model, the Bayesian approach compares the average fit of a model. This imposes a "natural" penalty for parameters, because each additional parameter introduces a dimension that must be averaged over. If that dimension introduces substantial parameter space with small likelihood, and little space that improves the likelihood, it will decrease the marginal likelihood. Thus, unlike the maximum likelihood, adding a parameter to a model can decrease the *marginal* likelihood, which ensures that more parameter-rich models are not automatically preferred.

The ratio of two marginal likelihoods gives us the factor by which the average fit of the model in the numerator is better or worse than the model in the denominator. This is called the Bayes factor (Jeffreys, 1935). We can again leverage Bayes' rule to gain more intuition for how marginal likelihoods and Bayes factors guide Bayesian model selection by writing it in terms of the posterior probability of a model, M_1, among N candidate models:

$$p(M_1 \mid D) = \frac{p(D \mid M_1)p(M_1)}{\sum_{i=1}^{N} p(D \mid M_i)p(M_i)}. \tag{3}$$

This shows us that the posterior probability of a model is proportional to the prior probability multiplied by the marginal likelihood of that model. Thus, the marginal likelihood is how the data update our prior beliefs about a model. As a result, it is often simply referred to as "the evidence" (MacKay, 2005). If we look at the ratio of the posterior probabilities of two models,

$$\frac{p(M_1 \mid D)}{p(M_2 \mid D)} = \frac{p(D \mid M_1)}{p(D \mid M_2)} \times \frac{p(M_1)}{p(M_2)}, \tag{4}$$

we see that the Bayes factor is the factor by which the prior odds of a model is multiplied to give use the posterior odds. Thus, marginal likelihoods and their ratios give us intuitive measures of how much the data "favor" one model over another, and these measures have natural probabilistic interpretations. However, marginal likelihoods and Bayes factors do not offer a panacea for model choice. As Equation 1 shows, weighting the average likelihood by the prior causes marginal likelihoods to be inherently sensitive to the prior distributions placed on the models' parameters. To gain more intuition about what this means and how Bayesian model choice differs from parameter estimation, let's use a simple, albeit contrived, example of flipping a coin.

2.1 A coin-flipping example

Let's assume we are interested in the probability of a coin we have not seen landing heads-side up when it is flipped (θ); we refer to this as the rate of landing heads up to avoid confusion with other uses of the word probability. Our plan is to flip this coin 100 times and count the number of times it lands heads up, which we model as a random outcome from a binomial distribution. Before flipping, we decide to compare four models that vary in our prior assumptions about the probability of the coin landing heads up (Figure 1): We assume

1. all values are equally probable (M_1: $\theta \sim \text{Beta}(1, 1)$),
2. the coin is likely weighted to land mostly "heads" or "tails" (M_2: $\theta \sim \text{Beta}(0.6, 0.6)$),
3. the coin is probably fair (M_3: $\theta \sim \text{Beta}(5.0, 5.0)$), and
4. the coin is weighted to land tails side up most of time (M_4: $\theta \sim \text{Beta}(1.0, 5.0)$).

We use beta distributions to represent our prior expectations, because the beta is a conjugate prior for the binomial likelihood function. This allows us to obtain the posterior distribution and marginal likelihood analytically.

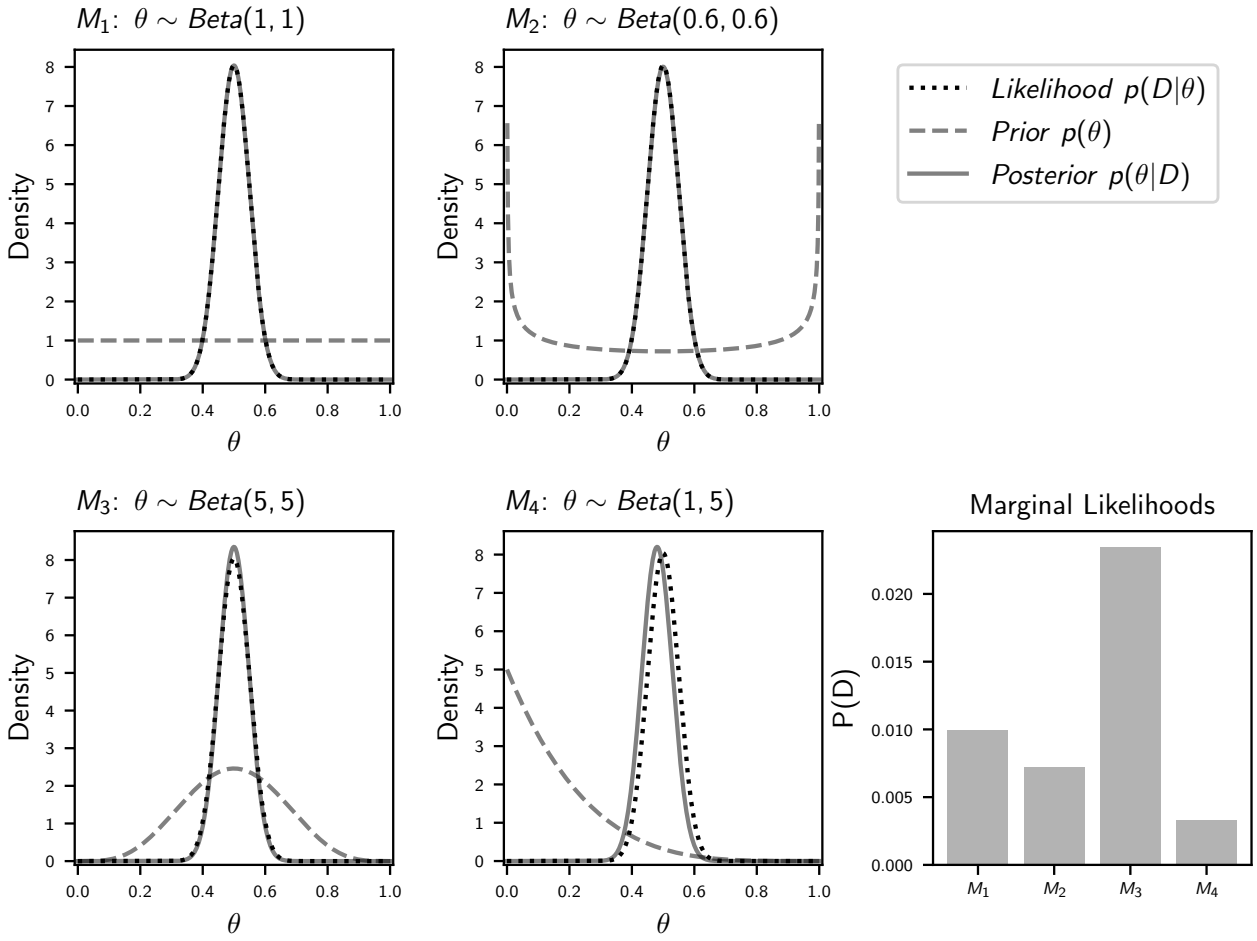

Figure 1. An illustration of the posterior probability densities and marginal likelihoods of the four different prior assumptions we made in our coin-flipping experiment. The data are 50 "heads" out of 100 coin flips, and the parameter, θ, is the probability of the coin landing heads side up. The binomial likelihood density function is proportional to a $\mathrm{Beta}(51, 51)$ and is the same across the four different beta priors on θ (M_1–M_4). The posterior of each model is a $\mathrm{Beta}(\alpha + 50, \beta + 50)$ distribution. The marginal likelihoods ($P(D)$; the average of the likelihood density curve weighted by the prior) of the four models are compared.

After flipping the coin and observing that it landed heads side up 50 times, we can calculate the posterior probability distribution for the rate of landing heads up under each of our four models:

$$p(\theta \mid D, M_i) = \frac{p(D \mid \theta, M_i)p(\theta \mid M_i)}{p(D \mid M_i)}. \tag{5}$$

Doing so, we see that regardless of our prior assumptions about the rate of the coin landing heads, the posterior distribution is very similar (Figure 1). This makes sense; given we observed 50 heads out of 100 flips, values for θ toward zero and one are extremely unlikely, and the posterior is dominated by the likelihood of values near 0.5.

Given the posterior distribution for θ is very robust to our prior assumptions, we might assume that each of our four models explain the data similarly well. However, to compare their ability to explain the data, we need to average (integrate) the likelihood density function over all possible values of θ, weighting by the prior:

$$p(D \mid M_i) = \int_\theta p(D \mid \theta, M_i)p(\theta \mid M_i)\mathrm{d}\theta. \tag{6}$$

Looking at the plots in Figure 1 we see that the models that place a lot of prior weight on values of θ that do not explain the data well (i.e., have small likelihood) have a much smaller marginal likelihood. Thus, even if we have very informative data that make the posterior distribution robust to prior assumptions, this example illustrates that the marginal likelihood of a model can still be very sensitive to the prior assumptions we make about the parameters.

Because of this inherent sensitivity to the priors, we have to take more care when choosing priors on the models' parameters when our goal is to compare models versus estimating parameters. For example, in Bayesian phylogenetics, it is commonplace to use "uninformative" priors, some of which are improper (i.e., they do not integrate to one). The example above demonstrates that if we have informative data, this objective Bayesian strategy (Jeffreys, 1961; Berger, 2006) is defensible if our goal is to infer the posterior distribution of a model;

we are hedging our bets against specifying a prior that concentrates its probability density outside of where the true value lies, and we can rely on the informative data to dominate the posterior. However, this strategy is much harder to justify if our goal is to compare marginal likelihoods among models. First of all, models with improper priors do not have a well-defined marginal likelihood and should not be used when comparing models (Baele et al., 2013b). Second, even if diffuse priors are proper, they could potentially sink the marginal likelihood of good models by placing excessive weight in biologically unrealistic regions of parameter space with low likelihood. Thus, if our goal is to leverage Bayesian model choice to learn about the processes that gave rise to our data, a different strategy is called for. One option is to take a more subjective Bayesian approach (Lad, 1996; Lindley, 2000; Goldstein, 2006) by carefully choosing prior distributions for the models' parameters based on existing knowledge. In the era of "big data," one could also use a portion of their data to inform the priors, and the rest of the data for inference. Alternatively, we can use hierarchical models that allow the data to inform the priors on the parameters (e.g., Suchard et al., 2003a).

We have developed an interactive version of Figure 1 where readers can vary the parameters of the coin-flip experiment and prior assumptions to further gain intuition for marginal likelihoods (https://kerrycobb.github.io/beta-binomial-web-demo/). It's worth noting that this pedagogical example is somewhat contrived given that the models we are comparing are simply different priors. Using the marginal likelihood to choose a prior is dubious, because the "best" prior will always be a point mass on the maximum likelihood estimate. Nonetheless, the principles of (and differences between) Bayesian parameter estimation and model choice that are illustrated by this example are directly relevant to more practical Bayesian inference settings. Now we turn to methods for approximating the marginal likelihood of phylogenetic models, where simple analytical solutions are generally not possible. Nonetheless, the same fundamental principles apply.

3 Methods for marginal likelihood approximation

For all but the simplest of models, the summation and integrals in Equation 1 are analytically intractable. This is particularly true for phylogenetic models, which have a complex structure containing both discrete and continuous elements. Thus, we must resort to numerical techniques to approximate the marginal likelihood.

Perhaps the simplest numerical approximation of the marginal likelihood is to draw samples of a model's parameters from their respective prior distributions. This turns the intractable integral into a sum of the samples' likelihoods. Because the prior weight of each sample is one in this case, the marginal likelihood can be approximated by simply calculating the average likelihood of the prior samples. Alternatively, if we have a sample of the parameters from the posterior distribution—like one obtained from a "standard" Bayesian phylogenetic analysis via MCMC—we can again use summation to approximate the integral. In this case, the weight of each sample is the ratio of the prior density to the posterior density. As a result, the sum simplifies to the harmonic mean (HM) of the likelihoods from the posterior sample (Newton and Raftery, 1994). Both of these techniques can be thought of as importance-sampling integral approximations. Whereas both provide unbiased estimates of the marginal likelihood in theory, they can suffer from very large Monte Carlo error due to the fact that the prior and posterior are often *very* divergent, with the latter usually *much* more peaked than the former due to the strong influence of the likelihood. A finite sample from the prior will often yield an underestimate of the marginal likelihood, because the region of parameter space with high likelihood is likely to be missed. In comparison, a finite sample from the posterior will almost always lead to an overestimate (Lartillot and Philippe, 2006; Xie et al., 2011; Fan et al., 2011), because it will contain too few samples outside of the region of high likelihood, where the prior weight "penalizes" the average likelihood. However, Baele et al. (2016) showed that for trees with 3–6 tips and relatively simple models, the average likelihood of a very large sample from the prior (30–50 billion samples) can yield accurate estimates of the marginal likelihood.

Recent methods developed to estimate marginal likelihoods generally fall into two categories for dealing with the sharp contrast between the prior and posterior that cripples the simple approaches mentioned above. One general strategy is to turn the giant leap between the unnormalized posterior and prior into many small steps across intermediate distributions; methods that fall into this category require samples collected from the intermediate distributions. The second strategy is to turn the giant leap between the posterior and prior into a smaller leap between the posterior and a reference distribution that is as similar as possible to the posterior; many methods in this category only require samples from the posterior distribution. These approaches are not mutually exclusive (e.g., see Fan et al. (2011)), but they serve as a useful way to categorize many of the methods available for approximating marginal likelihoods. In practical terms, the first strategy is computationally expensive, because samples need to be collected from each step between the posterior and prior, which is not normally part of a standard Bayesian phylogenetic analysis. The second strategy can be very inexpensive for methods that attempt to approximate the marginal likelihood using only the posterior samples collected from a typical analysis.

3.1 Approaches that bridge the prior and posterior with small steps

3.1.1 Path sampling (PS)

Lartillot and Philippe (2006) introduced path sampling (Gelman and Meng, 1998) (also called thermodynamic integration) to phylogenetics to address the problem that the posterior is often dominated by the likelihood and very divergent from the prior. Rather than restrict themselves to a sample from the posterior, they collected MCMC samples from a series of distributions between the prior and posterior. Specifically, samples are taken from a series of power-posterior distributions, $p(D \mid T, \theta, M)^{\beta} p(T, \theta \mid M)$, where the likelihood is raised to a power β. When $\beta = 1$, this is equal to the unnormalized joint posterior, which integrates to what we want to know, the marginal likelihood. When $\beta = 0$, this is equal to the joint prior distribution, which, assuming we are using proper prior probability distributions, integrates

to 1. If we integrate the power posterior expectation of the derivative with respect to β of the log power posterior over the interval (0–1) with respect to β, we get the log ratio of the normalizing constants when β equals 1 and 0, and since we know the constant is 1 when β is zero, we are left with the marginal likelihood. Lartillot and Philippe (2006) approximated this integral by summing over MCMC samples taken from a discrete number of β values evenly distributed between 1 and 0.

3.1.2 Stepping-stone (SS) sampling

The stepping-stone method introduced by Xie et al. (2011) is similar to PS in that it also uses samples from power posteriors, but the idea is not based on approximating the integral per se, but by the fact that we can accurately use importance sampling to approximate the ratio of normalizing constants with respect to two pre-chosen consecutive β values at each step between the posterior and prior. Also, Xie et al. (2011) chose the values of β for the series of power posteriors from which to sample so that most were close to the prior (reference) distribution, rather than evenly distributed between 0 and 1. This is beneficial, because most of the change happens near the prior; the likelihood begins to dominate quickly, even at small values of β. The stepping-stone method results in more accurate estimates of the marginal likelihood with fewer steps than PS (Xie et al., 2011).

3.1.3 Generalized stepping stone (GSS)

The most accurate estimator of marginal likelihoods available to date, the generalized stepping-stone (GSS) method, combines both strategies we are using to categorize methods by taking many small steps from a starting point (reference distribution) that is much closer to the posterior than the prior (Fan et al., 2011). Fan et al. (2011) improved upon the original stepping-stone method by using a reference distribution that, in most cases, will be much more similar to the posterior than the prior. The reference distribution has the same form as the joint prior, but each marginal prior distribution is adjusted so that its mean and

variance matches the corresponding sample mean and variance of an MCMC sample from the posterior. This guarantees that the support of the reference distribution will cover the posterior.

Initially, the application of the GSS method was limited, because it required that the topology be fixed, because there was no reference distribution across topologies. However, Holder et al. (2014) introduced such a distribution on trees, allowing the GSS to approximate the fully marginalized likelihood of phylogenetic models. Baele et al. (2016) introduced additional reference distributions on trees under coalescent models. Furthermore, Wu et al. (2014) and Rannala and Yang (2017) showed that the GSS and PS methods remain statistically consistent and unbiased when the topology is allowed to vary.

Based on intuition, it may seem that GSS would fail to adequately penalize the marginal likelihood, because it would lack samples from regions of parameter space with low likelihood (i.e., it does not use samples from the prior). However, importance sampling can be used to estimate the ratio of the normalizing constant of the posterior distribution (i.e., the marginal likelihood) to the reference distribution. As long as the reference distribution is proper, such that its normalizing constant is 1.0, this ratio is equal to the marginal likelihood. As a result, any proper reference distribution that covers the same parameter space as the posterior will work. The closer the reference is to the posterior, the easier it is to estimate the ratio of their normalizing constants (and thus the marginal likelihood). In fact, at the extreme that the reference distribution matches the posterior, we can determine the marginal likelihood exactly with only a single sample, because the difference in their densities is solely due to the normalizing constant of the posterior distribution (Fan et al., 2011).

All of the methods discussed below under “Approaches that use only posterior samples” are based on this idea of estimating the unknown normalizing constant of the posterior (the marginal likelihood) by "comparing" it to a reference distribution with a known normalizing constant (or at least a known difference in normalizing constant). What is different about GSS is the use of samples from a series of power-posterior distributions in between the ref-

erence and the posterior, which make estimating the ratio of normalizing constants between each sequential pair of distributions more accurate.

The fact that PS, SS, and GSS all use samples from a series of power-posterior distributions raises some important practical questions: How many power-posterior distributions are sufficient, how should they be spaced between the reference and posterior distribution, and how many MCMC samples are needed from each? There are no simple answers to these questions, because they will vary depending on the data and model. However, one general strategy that is clearly advantageous is having most of the β values near zero so that most of the power-posterior distributions are similar to the reference distribution (Lepage et al., 2007; Xie et al., 2011; Baele et al., 2016). Also, a practical approach to assess if the number of β values and the number of samples from each power posterior is sufficient is to estimate the marginal likelihood multiple times (starting with different seeds for the random number generator) for each model to get a measure of variance among estimates. It is difficult to quantify how much variance is too much, but the estimates for a model should probably be within a log likelihood unit or two from each other, and the ranking among models should be consistent. It can also be useful to check how much the variance among estimates decreases after repeating the analysis with more β values and/or more MCMC sampling from each step; a large decrease in variance suggests the sampling scheme was insufficient.

3.1.4 Sequential Monte Carlo (SMC)

Another approach that uses sequential importance-sampling steps is sequential Monte Carlo (SMC), also known as particle filtering (Gordon et al., 1993; Del Moral, 1996; Liu and Chen, 1998). Recently, SMC algorithms have been developed for approximating the posterior distribution of phylogenetic trees (Bouchard-Côté et al., 2012; Bouchard-Côté, 2014; Wang et al., 2018a). While inferring the posterior, SMC algorithms can approximate the marginal likelihood of the model "for free," by keeping a running average of the importance-sampling weights of the trees (particles) along the way. SMC algorithms hold a lot of promise for

complementing MCMC in Bayesian phylogenetics due to their sequential nature and ease with which the computations can be parallelized (Bouchard-Côté et al., 2012; Dinh et al., 2018; Fourment et al., 2018; Wang et al., 2018a). See Bouchard-Côté (2014) for an accessible treatment of SMC in phylogenetics.

Wang et al. (2018a) introduced a variant of SMC into phylogenetics that, similar to path sampling and stepping stone, transitions from a sample from the prior distribution to the posterior across a series of distributions where the likelihood is raised to a power (annealing). This approach provides an estimator of the marginal likelihood that is unbiased from both a statistical and computational perspective. Also, their approach maintains the full state space of the model while sampling across the power-posterior distributions, which allows them to use standard Metropolis-Hastings algorithms from the MCMC literature for the proposals used during the SMC. This should make the algorithm easier to implement in existing phylogenetic software compared to other SMC approaches that build up the state space of the model during the algorithm. Under the simulation conditions they explored, Wang et al. (2018a) showed that the annealed SMC algorithm compared favorably to MCMC and SS in terms of sampling the posterior distribution and estimating the marginal likelihood, respectively.

3.1.5 Nested sampling (NS)

Recently, Maturana R. et al. (2018) introduced the numerical technique known as nested sampling (Skilling, 2006) to Bayesian phylogenetics. This tries to simplify the multi-dimensional integral in Equation 1 into a one-dimensional integral over the cumulative distribution function of the likelihood. The latter can be numerically approximated using basic quadrature methods, essentially summing up the area of polygons under the likelihood function. The algorithm works by starting with a random sample of parameter values from the joint prior distribution and their associated likelihood scores. Sequentially, the sample with the lowest likelihood is removed and replaced by another random sample from the prior with the

constraint that its likelihood must be larger than the removed sample. The approximate marginal likelihood is a running sum of the likelihood of these removed samples with appropriate weights. Re-sampling these removed samples according to their weights yields a posterior sample at no extra computational cost. Initial assessment of NS suggest it performs similarly to GSS. As with SMC, NS seems like a promising complement to MCMC for both approximating the posterior and marginal likelihood of phylogenetic models.

3.2 Approaches that use only posterior samples

3.2.1 Generalized harmonic mean (GHM)

Gelfand and Dey (1994) introduced a generalized harmonic mean estimator that uses an arbitrary normalized reference distribution, as opposed to the prior distribution used in the HM estimator, to weight the samples from the posterior. If the chosen reference distribution is more similar to the posterior than the prior (i.e., a "smaller leap" as discussed above), the GHM estimator will perform better than the HM estimator. However, for high-dimensional phylogenetic models, choosing a suitable reference distribution is very challenging, especially for tree topologies. As a result, the GHM estimator has not been used for comparing phylogenetic models. However, recent advances on defining a reference distribution on trees (Holder et al., 2014; Baele et al., 2016) makes the GHM a tenable option in phylogenetics.

As discussed above, the HM estimator is unbiased in theory, but can suffer from very large Monte Carlo error in practice. The degree to which the GHM estimator solves this problem will depend on how much more similar the chosen reference distribution is to the posterior compared to prior. Knowing whether it is similar enough in practice will be difficult without comparing the estimates to other unbiased methods with much smaller Monte Carlo error (e.g., GSS, PS, or SMC).

3.2.2 Inflated-density ratio (IDR)

The inflated-density ratio estimator solves the problem of choosing a reference distribution by using a perturbation of the posterior density; essentially the posterior is "inflated" from the center by a known radius (Petris and Tardella, 2007; Arima and Tardella, 2012, 2014). As one might expect, the radius must be chosen carefully. The application of this method to phylogenetics has been limited by the fact that all parameters must be unbounded; any parameters that are bounded (e.g., must be positive) must be re-parameterized to span the real number line, perhaps using log transformation. As a result, this method cannot be applied directly to MCMC samples collected by popular Bayesian phylogenetic software packages. Nonetheless, the IDR estimator has recently been applied to phylogenetic models (Arima and Tardella, 2014), including in settings where the topology is allowed to vary (Wu et al., 2014). Initial applications of the IDR are very promising, demonstrating comparable accuracy to methods that sample from power-posterior distributions while avoiding such computation (Arima and Tardella, 2014; Wu et al., 2014). Currently, however, the IDR has only been used on relatively small datasets and simple models of character evolution. More work is necessary to determine whether the promising combination of accuracy and computational efficiency holds for large datasets and rich models.

3.2.3 Partition-weighted kernel (PWK)

Recently, Wang et al. (2018b) introduced the partition weighted kernel (PWK) method of approximating marginal likelihoods. This approach entails partitioning parameter space into regions within which the posterior density is relatively homogeneous. Given the complex structure of phylogenetic models, it is not obvious how this would be done. As of yet, this method has not been used for phylogenetic models. However, for simulations of mixtures of bivariate normal distributions, the PWK outperforms the IDR estimator (Wang et al., 2018b). Thus, the method holds promise if it can be adapted to phylogenetic models.

4 Uses of marginal likelihoods

The application of marginal likelihoods to compare phylogenetic models is rapidly gaining popularity. Rather than attempt to be comprehensive, below we highlight examples that represent some of the diversity of questions being asked and the insights that marginal likelihoods can provide about our data and the evolutionary processes giving rise to them.

4.1 Comparing partitioning schemes

One of the earliest applications of marginal likelihoods in phylogenetics was to choose among ways of assigning models of substitution to different subsets of aligned sites. This became important when phylogenetics moved beyond singe-locus trees to concatenated alignments of several loci. Mueller et al. (2004), Nylander et al. (2004), and Brandley et al. (2005) used Bayes factors calculated from harmonic mean estimates of marginal likelihoods to choose among different strategies for partitioning aligned characters to substitution models. All three studies found that the model with the most subsets was strongly preferred. Nylander et al. (2004) also showed that removing parameters for which the data seemed to have little influence decreased the HM estimates of the marginal likelihood, suggesting that the HM estimates might favor over-parameterized models. These findings could be an artefact of the tendency of the HM estimator to overestimate marginal likelihoods and thus underestimate the "penalty" associated with the prior weight of additional parameters. However, Brown and Lemmon (2007) showed that for simulated data, HM estimates of Bayes factors can have a low error rate of over-partitioning an alignment.

Fan et al. (2011) showed that, again, the HM estimator strongly favors the most partitioned model for a four-gene alignment from cicadas (12 subsets partitioned by gene and codon position). However, the marginal likelihoods estimated via the generalized stepping stone method favor a much simpler model (3 subsets partitioned by codon position). This demonstrates how the HM method fails to penalize the marginal likelihood for the weight of

the prior when applied to finite samples from the posterior. It also suggests that relatively few, well-assigned subsets can go a long way to explain the variation in substitution rates among sites.

Baele and Lemey (2013) compared the marginal likelihoods of alternative partitioning strategies (in combination with either strict or relaxed-clock models) for an alignment of whole mitochondrial genomes of carnivores. They used the harmonic mean, stabilized harmonic mean (Newton and Raftery, 1994), path sampling, and stepping-stone estimators. For all 41 models they evaluated, both harmonic mean estimators returned much larger marginal likelihoods than path sampling and stepping stone, again suggesting these estimators based solely on the posterior sample are unable to adequately penalize the models. They also found that by allowing the sharing of information among partitions via hierarchical modeling (Suchard et al., 2003a), the model with the largest PS and SS-estimated marginal likelihood switched from a codon model to a nucleotide model partitioned by codon position. This demonstrates the sensitivity of marginal likelihoods to prior assumptions.

4.2 Comparing models of character substitution

Lartillot and Philippe (2006) used path sampling to compare models of amino-acid substitution. They found that the harmonic mean estimator favored the most parameter rich model for all five datasets they explored, whereas the path-sampling estimates favored simpler models for three of the datasets. This again demonstrates that accurately estimated marginal likelihoods can indeed "penalize" for over-parameterization of phylogenetic models. More importantly, this work also revealed that modeling heterogeneity in amino acid composition across sites of an alignment better explains the variation in biological data.

4.3 Comparing "relaxed clock" models

Lepage et al. (2007) used path sampling to approximate Bayes factors comparing various "relaxed-clock" phylogenetic models for three empirical datasets. They found that models

in which the rate of substitution evolves across the tree (autocorrelated rate models) better explain the empirical sequence alignments they investigated than models that assume the rate of substitution on each branch is independent (uncorrelated rate models). This provides insight into how the rate of evolution evolves through time.

Baele et al. (2013b) demonstrated that modeling among-branch rate variation with a lognormal distribution tends to explain mammalian sequence alignments better than using an exponential distribution. They used marginal likelihoods (PS and SS estimates) and Bayesian model averaging to compare the fit of lognormally and exponentially distributed priors on branch-specific rates of nucleotide substitution (i.e., relaxed clocks) for almost 1,000 loci from 12 mammalian species. They found that the lognormal relaxed-clock was a better fit for almost 88% of the loci. Baele et al. (2012) also used marginal likelihoods to demonstrate the importance of using sampling dates when estimating time-calibrated phylogenetic trees. They used path-sampling and stepping-stone methods to estimate the marginal likelihoods of strict and relaxed-clock models for sequence data of herpes viruses. They found that when the dates the viruses were sampled were provided, a strict molecular clock was the best fit model, but when the dates were excluded, relaxed-clock models were strongly favored. Their findings show that using information about the ages of the tips can be critical for accurately modeling processes of evolution and inferring evolutionary history.

4.4 Comparing demographic models

Baele et al. (2012) used the path-sampling and stepping-stone estimators for marginal likelihoods to compare the fit of various demographic models to the HIV-1 group M data of Worobey et al. (2008), and Methicillin-resistant *Staphylococcus aureus* (MRSA) data of Gray et al. (2011). They found that a flexible, nonparametric model that enforces no particular demographic history is a better explanation of the HIV and MRSA sequence data than exponential and logistic population growth models. This suggests that traditional parametric growth models are not the best predictors of viral and bacterial epidemics.

4.5 Measuring phylogenetic information content across genomic data sets

Not only can we use marginal likelihoods to learn about evolutionary models, but we can also use them to learn important lessons about our data. Brown and Thomson (2017) explored six different genomic data sets that were collected to infer phylogenetic relationships within Amniota. For each locus across all six data sets, they used the stepping-stone method (Xie et al., 2011) to approximate the marginal likelihood of models that included or excluded a particular branch (bipartition) in the amniote tree. This allowed Brown and Thomson (2017) to calculate, for each gene, Bayes factors as measures of support for or against particular relationships, some of which are uncontroversial (e.g., the monophyly of birds) and others contentious (e.g., the placement of turtles).

Such use of marginal likelihoods to compare topologies, or constraints on topologies, raises some interesting questions. Bergsten et al. (2013) showed that using Bayes factors for topological tests can result in strong support for a constrained topology over an unconstrained model for reasons other than the data supporting the branch (bipartition) being constrained. This occurs when the data support other branches in the tree that make the constrained branch more likely to be present just by chance, compared to a diffuse prior on topologies. This is not a problem with the marginal likelihoods (or their estimates), but rather how we interpret the results of the Bayes factors; if we want to interpret it as support for a particular relationship, we have to be cognizant of the topology space we are summing over under both models. Brown and Thomson (2017) tried to limit the effect of this issue by constraining all "uncontroversial" bipartitions when they calculate the marginal likelihoods of models with and without a particular branch, essentially enforcing an informative prior across topologies under both models.

Brown and Thomson's 2017 use of marginal likelihoods allowed them to reveal a large degree of variation among loci in support for and against relationships that was masked by the corresponding posterior probabilities estimated by MCMC. Furthermore, they found that a

small number of loci can have a large effect on the tree and associated posterior probabilities of branches inferred from the combined data. For example, they showed that including or excluding just two loci out of the 248 locus dataset of (Chiari et al., 2012) resulted in a posterior probability of 1.0 in support of turtles either being sister to crocodylians or archosaurs (crocodylians and birds), respectively. By using marginal likelihoods of different topologies, Brown and Thomson (2017) were able to identify these two loci as putative paralogs due to their strikingly strong support for turtles being sister to crocodylians. This work demonstrates how marginal likelihoods can simultaneously be used as a powerful means of controlling the quality of data in "phylogenomics", while informing us about the evolutionary processes that gave rise to our data.

Furthermore, Brown and Thomson (2017) found that the properties of loci commonly used as proxies for the reliability of phylogenetic signal (rate of substitution, how "clock-like" the rate is, base composition heterogeneity, amount of missing data, and alignment uncertainty) were poor predictors of Bayes factor support for well-established amniote relationships. This suggests these popular rules of thumb are not useful for identifying "good" loci for phylogenetic inference.

4.6 Phylogenetic factor analysis

The goal of comparative biology is to understand the relationships among a potentially large number of phenotypic traits across organisms. To do so correctly, we need to account for the inherent shared ancestry underlying all life(Felsenstein, 1985). A lot of progress has been made for inferring the relationship between pairs of phenotypic traits as they evolve across a phylogeny, but a general and efficient solution for large numbers of continuous and discrete traits has remained elusive. Tolkoff et al. (2018) introduced Bayesian factor analysis to a phylogenetic framework as a potential solution. Phylogenetic factor analysis works by modeling a small number of unobserved (latent) factors that evolve independently across the tree, which give rise to the large number of observed continuous and discrete phenotypic

traits. This allows correlations among traits to be estimated, without having to model every trait as a conditionally independent process.

The question that immediately arises is, what number of factors best explains the evolution of the observed traits? To address this, Tolkoff et al. (2018) use path sampling to approximate the marginal likelihood of models with different numbers of traits. To do so, they extend the path sampling method to handle the latent variables underlying the discrete traits by softening the thresholds that delimit the discrete character states across the series of power posteriors. This new approach leverages Bayesian model comparison via marginal likelihoods to learn about the processes governing the evolution of multidimensional phenotypes.

4.7 Comparing phylogeographic models

Phylogeographers are interested in explaining the genetic variation within and among species across a landscape. As a result, we are often interested in comparing models that include various combinations of micro and macro-evolutionary processes and geographic and ecological parameters. Deriving the likelihood function for such models is often difficult and, as a result, model choice approaches that use approximate-likelihood Bayesian computation (ABC) are often used.

At the forefront of generalizing phylogeographic models is an approach that is referred to as iDDC, which stands for integrating distributional, demographic, and coalescent models (Papadopoulou and Knowles, 2016). This approach simulates data under various phylogeographical models upon proxies for habitat suitability derived from species distribution models. To choose the model the best explains the empirical data, this approach uses the marginal densities of the models estimated via the ABC-GLM method and p-values derived from these densities (He et al., 2013) (Massatti and Knowles, 2016) (Bemmels et al., 2016) (Knowles and Massatti, 2017) (Papadopoulou and Knowles, 2016). This approach is an important step forward for bringing more biological realism into phylogeographical models.

However, our findings below (see section on "approximate-likelihood approaches" below) show that the marginal GLM density fitted to a truncated region of parameter space should not be interpreted as a marginal likelihood of the full model. Thus, these methods should be seen as a useful exploration of data, rather than rigorous hypothesis tests. Because ABC-GLM marginal densities fail to penalize parameters for their prior weight in regions of low likelihood, these approaches will likely be biased toward over-parameterized phylogeographical models. Nonetheless, knowledge of this bias can help guide interpretations of results.

4.8 Species delimitation

Calculating the marginal probability of sequence alignments (Grummer et al., 2013) and single-nucleotide polymorphisms (Leaché et al., 2014) under various multi-species coalescent models has been used to estimate species boundaries. By comparing the marginal likelihoods of models that differ in how they assign individual organisms to species, systematists can calculate Bayes factors to determine how much the genetic data support different delimitations. Using simulated data, Grummer et al. (2013) found that marginal likelihoods calculated using path sampling and stepping-stone methods outperformed harmonic mean estimators at identifying the true species delimitation model. Marginal likelihoods seem better able to distinguish some species delimitation models than others. For example, models that lump species together or reassign samples into different species produce larger marginal likelihood differences versus models that split populations apart (Grummer et al., 2013; Leaché et al., 2014). Current implementations of the multi-species coalescent assume strict models of genetic isolation, and oversplitting populations that exchange genes creates a difficult Bayesian model comparison problem that does not include the correct model (Leaché et al., 2018a,b).

Species delimitation using marginal likelihoods in conjunction with Bayes factors has some advantages over alternative approaches. The flexibility of being able to compare non-nested models that contain different numbers of species, or different species assignments, is one key advantage. The methods also integrate over gene trees, species trees, and other

model parameters, allowing the marginal likelihoods of delimitations to be compared without conditioning on any parameters being known. Marginal likelihoods also provide a natural way to rank competing models while automatically accounting for model complexity (Baele et al., 2012). Finally, it is unnecessary to assign prior probabilities to the alternative species delimitation models being compared. The marginal likelihood of a delimitation provides the factor by which the data update our prior expectations, regardless of what that expectation is (Equation 3). As multi-species coalescent models continue to advance, using the marginal likelihoods of delimitations will continue to be a powerful approach to learning about biodiversity.

5 Alternatives to marginal likelihoods for Bayesian model choice

5.1 Bayesian model averaging

Bayesian model averaging provides a way to avoid model choice altogether. Rather than infer the parameter of interest (e.g., the topology) under a single "best" model, we can incorporate uncertainty by averaging the posterior over alternative models. In situations where model choice is not the primary goal, and the parameter of interest is sensitive to which model is used, model averaging is arguably the best solution from a Bayesian standpoint. Nonetheless, when we jointly sample the posterior across competing models, we can use the posterior sample for the purposes of model choice. The frequency of samples from each model approximates its posterior probability, which can be used to approximate Bayes factors among models. Note, this approach is still based on marginal likelihoods—the marginal likelihood is how the data inform the model posterior probabilities, and the Bayes factor is simply a ratio of marginal likelihoods (Equations 3 & 4). However, by sampling across models, we can avoid calculating the marginal likelihoods directly.

Algorithms for sampling across models include reversible-jump MCMC (Green, 1995), Gibbs sampling (Neal, 2000), Bayesian stochastic search variable selection (George and McCulloch, 1993; Kuo and Mallick, 1998), and approximations of reversible-jump (Jones et al., 2015). In fact, the first application of Bayes factors for phylogenetic model comparison was performed by Suchard et al. (2001) via reversible-jump MCMC. This technique was also used in Bayesian tests of phylogenetic incongruence/recombination (Suchard et al., 2003b; Minin et al., 2005). In terms of selecting the correct "relaxed-clock" model from simulated data, Baele et al. (2013b) and Baele and Lemey (2014) showed that model-averaging performed similarly to the path-sampling and stepping-stone marginal likelihood estimators.

There are a couple of limitations for these approaches. First, a Bayes factor that includes a model with small posterior probability will suffer from Monte Carlo error. For example, unless a very large sample from the posterior is collected, some models might not be sampled at all. A potential solution to this problem is adjusting the prior probabilities of the models such that none of their posterior probabilities are very small (Carlin and Chib, 1995; Suchard et al., 2005). Second, and perhaps more importantly, for these numerical algorithms to be able to "jump" among models, the models being sampled need to be similar. Whereas the first limitation is specific to using model averaging to estimate Bayes factors, the second problem is more general.

In comparison, with estimates of marginal likelihooods in hand, we can compare any models, regardless of how different they are in terms of parameterization or relative probability. Alternatively, Lartillot and Philippe (2006) introduced a method of using path sampling to directly approximate the Bayes factor between two models that can be highly dissimilar. Similarly, Baele et al. (2013a) extended the stepping-stone approach of Xie et al. (2011) to do the same.

5.2 Measures of predictive performance

Another, albeit not an unrelated, way to compare models is based on their predictive power, with the idea that we should prefer the model that best predicts future data. There are many approaches to do this, but they are all centered around measuring the predictive power of a model using the marginal probability of new data (D') given our original data (D),

$$p(D' \mid M, D) = \sum_{T} \int_{\theta} p(D' \mid T, \theta, M) p(T, \theta \mid M, D) \mathrm{d}\theta, \tag{7}$$

which we will call the marginal posterior predictive likelihood. This has clear parallels to the marginal likelihood (see Equation 1), with one key difference: We condition on our knowledge of the original data, so that the average of the likelihood of the new data is now weighted by the *posterior* distribution rather than the prior. Thus, in situations where our data are informative and dominate the posterior distribution under each model, the marginal posterior predictive likelihood should be much less sensitive than the marginal likelihood to the prior distributions used for the models' parameters.

Whether one should favor a posterior-predictive perspective or marginal likelihoods will depend on the goals of a particular model-choice exercise and whether the prior is the appropriate penalty for adding parameters to a model. Regardless, posterior predictive measures of model fit are a valuable complement to marginal likelihoods. Methods based on the marginal posterior predictive likelihood tend to be labeled with one of two names depending on the surrogate they use for the "new" data (D'): (1) **cross-validation methods** partition the data under study into a training (D) and testing (D') dataset, whereas (2) **posterior-predictive methods** generate D' via simulation.

5.2.1 Cross-validation methods

With joint samples of parameter values from the posterior (conditional on the training data D), we can easily get a Monte Carlo approximation of the marginal posterior predictive likelihood (Equation 7) by simply taking the average probability of the testing data across the posterior samples of parameter values:

$$p(D' \mid M, D) \simeq \frac{1}{n} \sum_{i-1}^{n} p(D' \mid T_i, \theta_i, M), \tag{8}$$

where n is the number of samples from the posterior under model M. Lartillot et al. (2007) used this approach to show that a mixture model that accommodates among-site heterogeneity in amino acid frequencies is a better predictor of animal sequence alignments than standard amino-acid models. This corroborated the findings of Lartillot and Philippe (2006) based on path-sampling estimates of marginal likelihoods. Lewis et al. (2014) introduced a leave-one-out cross-validation approach to phylogenetics called the conditional predictive ordinates (CPO) method (Geisser, 1980; Gelfand et al., 1992; Chen et al., 2000). This method leaves one site out of the alignment to serve as the testing data to estimate $p(D' \mid M, D)$, which is equal to the posterior harmonic mean of the site likelihood (Chen et al., 2000). Summing the log of this value across all sites yields what is called the log pseudomarginal likelihood (LPML). Lewis et al. (2014) compared the estimated LPML to stepping-stone estimates of the marginal likelihood for selecting among models that differed in how they partitioned sites across a concatenated alignment of four genes from algae. The LPML favored a 12-subset model (partitioned by gene and codon position) as opposed to the 3-subset model (partitioned by codon) preferred by marginal likelihoods. This difference could reflect the lesser penalty against additional parameters imposed by the weight of the posterior (Equation 7) versus the prior (Equation 1).

5.2.2 Posterior-predictive methods

Alternatively, we can take a different Monte Carlo approach to Equation 7 and sample from $p(D' \mid M, D)$ by simulating datasets. For each posterior sample of the parameter values (conditional on all the data under study) we can simply simulate a new dataset based on those parameter values. We can then compare the observed data (D) to the sample of simulated datasets (D') from the posterior predictive distribution. In all but the most trivial phylogenetic datasets, it is not practical to compare the counts of site patterns directly, because there are too many possible patterns (e.g., four raised to the power of the number of tips for DNA data). Thus, we have to tolerate some loss of information by summarizing the data in some way to reduce the dimensionality. Once a summary statistic is chosen, perhaps the simplest way to evaluate the fit of the model is to approximate the posterior predictive p-value by finding the percentile of the statistic from the observed data out of the values of the statistic calculated from the simulated datasets (Rubin, 1984; Gelfand et al., 1992). Bollback (2002) explored this approach for phylogenetic models using simulated data, and found that a simple JC69 model (Jukes and Cantor, 1969) was often rejected for data simulated under more complex K2P (Kimura, 1980) and GTR (Tavaré et al., 1997) models. Lartillot et al. (2007) also used this approach to corroborate their findings based on marginal likelihoods (Lartillot and Philippe, 2006) and cross validation that allowing among-site variation in amino acid composition (i.e., the CAT model) leads to a better fit.

One drawback of the posterior predictive p-value is that it rewards models with large posterior predictive variance (Lewis et al., 2014). In other words, a model that produces a broad enough distribution of datasets can avoid the observed data falling into one of the tails. The method of Gelfand and Ghosh (1998) (GG) attempts to solve this problem by balancing the tradeoff between posterior predictive variance and goodness-of-fit. Lewis et al. (2014) introduced the GG method into phylogenetics and compared it to cross-validation (LPML) and stepping-stone estimates of marginal likelihoods for selecting among models that differed in how they partitioned the sites of a four-gene alignment of algae. Similar

to LPML, the GG method preferred the model with most subsets (12; partitioned by gene and codon position), in contrast to the marginal likelihood estimates, which favored the model partitioned by codon position (three subsets). Again, this difference could be due to the lesser penalty against parameters imposed by the weight of the posterior (Equation 7) versus the prior (Equation 1).

5.3 Approximate-likelihood approaches

Approximate-likelihood Bayesian computation (ABC) approaches (Tavaré et al., 1997; Beaumont et al., 2002) have become popular in situations where it is not possible (or undesirable) to derive and compute the likelihood function of a model. The basic idea is simple: by generating simulations under the model, the fraction of times that we generate a simulated dataset that matches the observed data is a Monte Carlo approximation of the likelihood. Because simulating the observed data exactly is often not possible (or extremely unlikely), simulations "close enough" to the observed data are counted, and usually a set of insufficient summary statistics are used in place of the data. Whether a simulated dataset is "close enough" to count is formalized as whether or not it falls within a zone of tolerance around the empirical data.

This simple approach assumes the likelihood within the zone of tolerance is constant. However, this zone usually needs to be quite large for computational tractability, so this assumption does not hold. Leuenberger and Wegmann (Leuenberger and Wegmann, 2010) proposed fitting a general linear model (GLM) to approximate the likelihood within the zone of tolerance. With the GLM in hand, the marginal likelihood of the model can simply be approximated by the marginal density of the GLM.

The accuracy of this estimator has not been assessed. However, there are good theoretical reasons to be skeptical of its accuracy. Because the GLM is only fit within the zone of tolerance (also called the "truncated prior"), it cannot account for the weight of the prior on the marginal likelihood outside of this region. Whereas the posterior distribution usually

is not strongly influenced by regions of parameter space with low likelihood, the marginal likelihood very much is. By not accounting for prior weight in regions of parameter space outside the zone of tolerance, where the likelihood is low, we predict this method will not properly penalize models and tend to favor models with more parameters.

To test this prediction, we assessed the behavior of the ABC-GLM method on 100 datasets simulated under the simplest possible phylogenetic model: two DNA sequences separated by a single branch along which the sequence evolved under a Jukes-Cantor model of nucleotide substitution (Jukes and Cantor, 1969). The simulated sequences were 10,000 nucleotides long, and the prior on the only parameter in the model, the length of the branch, was a uniform distribution from 0.0001 to 0.1 substitutions per site. For such a simple model, we used quadrature integration to calculate the marginal likelihood for each simulated alignment of two sequences. Integration using 1,000 and 10,000 steps and rectangular and trapezoidal quadrature rules all yielded identical values for the log marginal likelihood to at least five decimal places for all 100 simulated data sets, providing a very precise proxy for the true values. We used a sufficient summary statistic, the proportion of variable sites, for ABC analyses. However, the ABC-GLM and quadrature marginal likelihoods are not directly comparable, because the marginal probability of the proportion of variable sites versus the site pattern counts will be on different scales that are data set dependent. So, we compare the ratio of marginal likelihoods (i.e., Bayes factors) comparing the correct branch-length model [branch length $\sim$ uniform(0.0001, 0.1)] to a model with a prior approximately twice as broad [branch length $\sim$ uniform(0.0001, 0.2)]. As we noted in our coin-flipping example, using marginal likelihoods to compare priors is dubious, and we do not advocate selecting priors in this way. However, in this case, comparing the marginal likelihood under these two priors is useful, because it allows us to directly test our prediction that the ABC-GLM method will not be able to correctly penalize the marginal likelihood for the additional parameter space under the broader prior.

This very simple model is a good test of the ABC-GLM marginal likelihood estimator

for several reasons. The use of a sufficient statistic for a finite, one-dimensional model makes ABC nearly equivalent to a full-likelihood Bayesian method (Figure A1). Thus, this is a "best-case scenario" for the ABC-GLM approach. Also, we can use quadrature integration for very good proxies for the true Bayes factors. Lastly, the simple scenario gives us some analytical expectations for the behavior of ABC-GLM. If it cannot penalize the marginal likelihood for the additional branch length space in the model with the broader prior, the Bayes factor should be off by a factor of approximately 2, or more precisely $(0.2-0.0001)/(0.1-0.0001)$. As shown in Figure 2, this is exactly what we find. This confirms our prediction that the ABC-GLM approach cannot average over regions of parameter space with low likelihood and thus will be biased toward favoring models with more parameters. Given that the GLM approximation of the likelihood is only fit within a subset of parameter space with high likelihood, which is usually a *very* small region of a model, the marginal of the GLM should not be considered a marginal likelihood of the model. We want to emphasize that our findings in no way detract from the usefulness of ABC-GLM for parameter estimation.

Full details of these analyses, which were all designed atop the DendroPy phylogenetic API (version 4.3.0 commit 72ce015) (Sukumaran and Holder, 2010), can be found in Appendix A, and all of the code to replicate our results is freely available at `https://github.com/phyletica/abc-glm-marginal-test`.

6 Discussion

6.1 Promising future directions

As Bayesian phylogenetics continues to explore more complex models of evolution, and datasets continue to get larger, accurate and efficient methods of estimating marginal likelihoods will become increasingly important. Thanks to substantial work in recent years, robust methods have been developed, such as the generalized stepping-stone approach (Fan

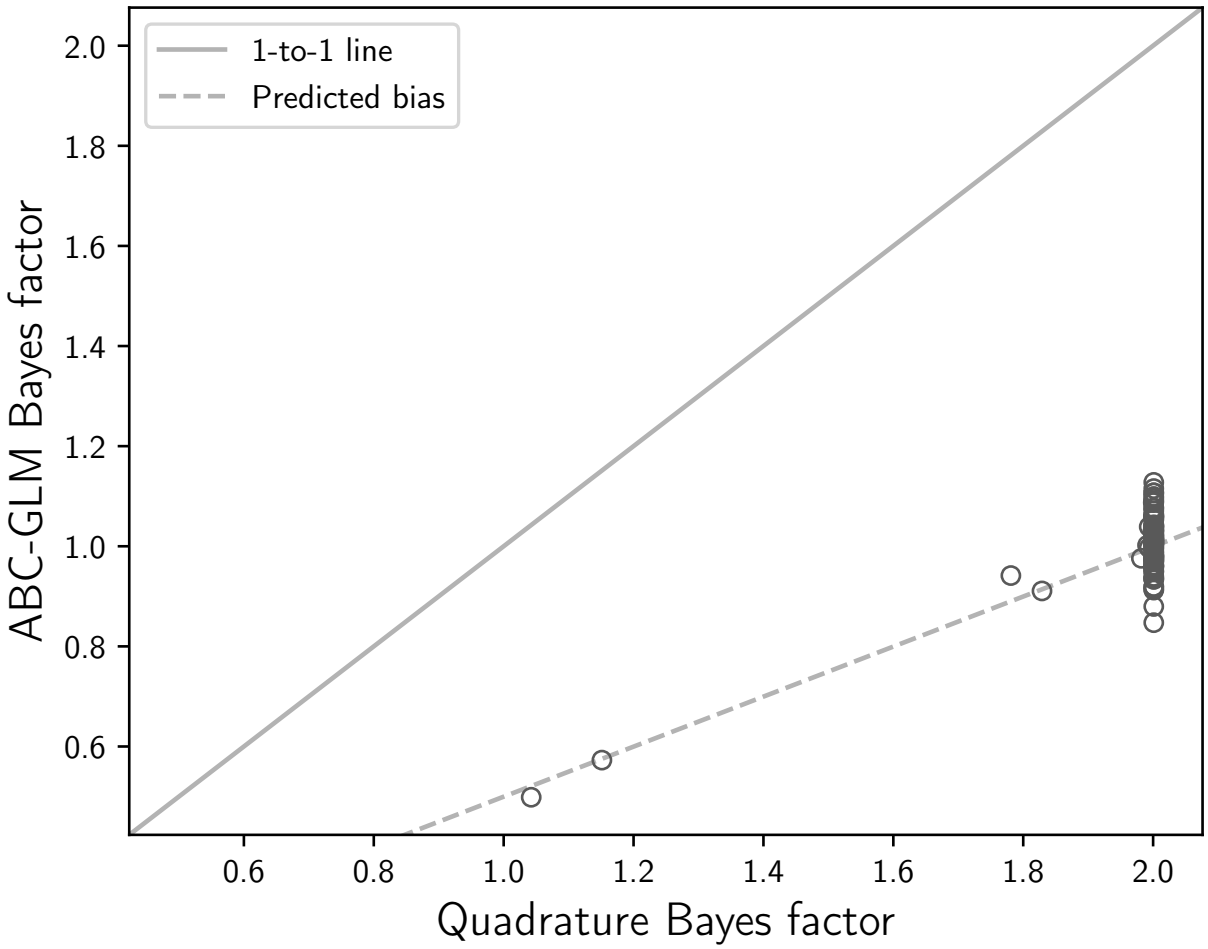

Figure 2. A comparison of the approximate-likelihood Bayesian computation general linear model (ABC-GLM) estimator of the marginal likelihood (Leuenberger and Wegmann, 2010) to quadrature integration approximations (Xie et al., 2011) for 100 simulated datasets. We compared the ratio of the marginal likelihood (Bayes factor) comparing the correct branch-length model [branch length $\sim$ uniform(0.0001, 0.1)] to a model with a broader prior on the branch length [branch length $\sim$ uniform(0.0001, 0.2)]. The solid line represents perfect performance of the ABC-GLM estimator (i.e., matching the "true" value of the Bayes factor). The dashed line represents the expected Bayes factor when failing to penalize for the extra parameter space (branch length 0.1 to 0.2) with essentially zero likelihood. Quadrature integration with 1,000 and 10,000 steps using the rectangular and trapezoidal rule produced identical values of log marginal likelihoods to at least five decimal places for all 100 simulated datasets.

et al., 2011). However, these methods are computationally demanding as they have to sample likelihoods across a series of power-posterior distributions that are not useful for parameter estimation. Recent work has introduced promising methods to estimate marginal likelihoods solely from samples from the posterior distribution. However, these methods remain difficult to apply to phylogenetic models, and their performance on rich models and large datasets remains to be explored.

Promising avenues for future research on methods for estimating marginal likelihoods of phylogenetic models include continued work on reference distributions that are as similar to the posterior as possible, but easy to formulate and use. This would improve the performance and applicability of the GSS and derivations of the GHM approach. Currently, the most promising method that works solely from a posterior sample is IDR. Making this method easier to apply to phylogenetic models and implementing it in popular Bayesian phylogenetic software packages, like RevBayes (Höhna et al., 2016) and BEAST (Suchard et al., 2018; Bouckaert et al., 2014) would be very useful, though nontrivial.

Furthermore, nested sampling and sequential Monte Carlo are exciting numerical approaches to Bayesian phylogenetics. These methods essentially use the same amount of computation to both sample from the posterior distribution of phylogenetic models and provide an approximation of the marginal likelihood. Both approaches are relatively new to phylogenetics, but hold a lot of promise for Bayesian phylogenetics generally and model comparison via marginal likelihoods specifically.

6.2 A fundamental challenge of Bayesian model choice

While the computational challenges to approximating marginal likelihoods are very real and will provide fertile ground for future research, it is often easy to forget about a fundamental challenge of Bayesian model choice. This challenge becomes apparent when we reflect on the differences between Bayesian model choice and parameter estimation. The posterior distribution of a model, and associated parameter estimates, are informed by the likelihood

function (Equation 2), whereas the posterior probability of that model is informed by the *marginal* likelihood (Equation 3). When we have informative data, the posterior distribution is dominated by the likelihood, and as a result our parameter estimates are often robust to prior assumptions we make about the parameters. However, when comparing models, we need to assess their overall ability to predict the data, which entails averaging over the entire parameter space of the model, not just the regions of high likelihood. As a result, marginal likelihoods and associated model choices can be very sensitive to priors on the *parameters* of each model, even when the data are very informative (Figure 1). This sensitivity to prior assumptions about parameters is inherent to Bayesian model choice based on marginal likelihoods (i.e., Bayes factors and Bayesian model averaging). However, other Bayesian model selection approaches, such as cross-validation and posterior-predictive methods, will be less sensitive to prior assumptions. Regardless, the results of any application of Bayesian model selection should be accompanied by an assessment of the sensitivity of those results to the priors placed on the models' parameters.

6.3 Conclusions

Marginal likelihoods are intuitive measures of model fit that are grounded in probability theory. As a result, they provide us with a coherent way of gaining a better understanding about how evolution proceeds as we accrue biological data. We highlighted how marginal likelihoods of phylogenetic models can be used to learn about evolutionary processes and how our data inform our models. Because shared ancestry is a fundamental property of life, the use of marginal likelihoods of phylogenetic models promises to continue to advance biology.

7 Funding

This work was supported by the National Science Foundation (grant numbers DBI 1308885 and DEB 1656004 to JRO).

8 Acknowledgments

We thank Mark Holder for helpful discussions about comparing approximate and full marginal likelihoods. We also thank Ziheng Yang and members of the Phyletica Lab (the phyleticians) for helpful comments that improved an early draft of this paper. We are grateful to Guy Baele, Nicolas Lartillot, Paul Lewis, two anonymous referees, and Associate Editor, Olivier Gascuel, for constructive reviews that greatly improved this work. The computational work was made possible by the Auburn University (AU) Hopper Cluster supported by the AU Office of Information Technology. This paper is contribution number 880 of the Auburn University Museum of Natural History.

Appendix A Methods for assessing performance of ABC-GLM estimator

We set up a simple scenario for assessing the performance of the method for estimating marginal likelihoods based on approximating the likelihood function with a general linear model (GLM) fitted to posterior samples collected via approximate-likelihoood Bayesian computation (ABC) (Leuenberger and Wegmann, 2010); hereforth referred to as ABC-GLM. The scenario is a DNA sequence, 10,000-nucleotides in length, that evolves along a branch according to a Jukes-Cantor continuous-time Markov chain (CTMC) model of nucleotide substitution (Jukes and Cantor, 1969). Because the Jukes-Cantor model forces the relative rates of change among the four nucleotides and the equilibrium nucleotide frequencies to be equal, there is only a single parameter in the model, the length of the branch, and the

direction of evolution along the branch does not matter.

A.1 Simulating data sets

We simulated 100 data sets under this model by

1. drawing 10,000 nucleotides of the "ancestral" sequence from their equilibrium frequencies ($\frac{1}{4}$),
2. drawing a branch length $\sim$ uniform(0.0001, 0.1), and
3. evolving the sequence along the branch according to the Jukes-Cantor CTMC model to get the "descendant" sequence.

This was done using the DendroPy phylogenetic API (version 4.3.0 commit 72ce015) (Sukumaran and Holder, 2010).

A.2 Calculating "true" Bayes factors

For each data set, we used quadrature approaches to approximate the marginal likelihood by integrating the posterior density over the branch length prior. We did this for two models:

1. the correct model [branch length $\sim$ uniform(0.0001, 0.1)], and
2. a model with a branch length prior slightly more than twice as broad [branch length $\sim$ uniform(0.0001, 0.2)], which we refer to as the "vague model".

For both models and for each dataset we used the rectangular and trapezoidal quadrature rules with 1,000 and 10,000 steps (i.e., four approximations of the marginal likelihood for each data set under each model). Across all 100 data sets and both models, all four approximations were identical to at least five decimal places. For each data set, we calculated the log Bayes factor comparing the correct model to the vague model.

A.3 Approximate-likelihood Bayesian computation

To collect an approximate posterior sample from the correct model for a data set, we first calculated the proportion of variable sites ($Pvar$) between the two sequences. Next, we simulated 50,000 datasets under the correct model, calculated $Pvar$ for each of them, and retained the 1,000 samples with the values of $Pvar$ closest to that calculated from the data. Lastly, we used ABCtoolbox version 1.1 Wegmann et al. (2010) to fit a GLM to the retained samples and calculate the marginal density of the GLM, using a bandwidth of 0.002. We did the same to obtain an ABC-GLM estimate of the marginal density for the vague model with two differences: (1) we drew the branch length for each prior sample from the vague prior [branch length $\sim$ uniform(0.0001, 0.2)], and (2) to maintain the same expected tolerance under both models, we simulated 100,000 datasets under the vague model (retaining the 1,000 samples closest to the $Pvar$ of the data).

For each data set, we calculated the log Bayes factor from the GLM marginal densities of the correct and vague model, and compared the ABC-GLM-estimated Bayes factor to the "true" Bayes factor calculated via quadrature integration (Figure 2).

A.4 Full-likelihood Markov chain Monte Carlo analyses

One goal of the simplicity of the above model is that the additional approximation of the ABC approach would be limited. All numerical Bayesian analyses, based on full or approximate likelihoods, suffer from Monte Carlo error associated with approximating the posterior with a finite number of samples. Approximate-likelihood methods usually suffer from two additional sources of approximation: (1) the full data are replaced with insufficient summary statistics, and (2) samples are retained that do not exactly match the data or summary statistics (i.e., the "tolerance" of ABC). In our analyses described above, we avoided the former source of error by using a sufficient statistic. We hoped to minimize the latter source of error by evaluating many samples from a one-dimensional model with finite bounds; we also kept this source of error approximately equal for both models by sampling

in proportion to the width of the model.

To verify that the error introduced by the tolerance of the ABC analyses was minimal, we compared the branch length estimates to those estimated by full-likelihood Markov chain Monte Carlo (MCMC). For each data set, under both models, we ran a chain for 10,000 generations, sampling every 10 generations. All chains appeared to reach stationarity by the first sample (10th generation). We plotted the branch length estimated via ABC-GLM and MCMC under both the true and vague models against the true branch lengths. The results of all four analyses across all 100 data sets are almost indistinguishable (Figure A1), confirming that the approximation introduced by the tolerance is very minimal. Our ABC-GLM analyses are essentially equivalent to full-likelihood Bayesian analyses, creating a "best-case scenario" for evaluating the marginal likelihood estimates of the ABC-GLM method.

A.5 Reproducibility

All of the code to replicate our results is freely available at https://github.com/phyletica/abc-glm-marginal-test.

References

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716–723.

Arima, S. and L. Tardella. 2012. Improved harmonic mean estimator for phylogenetic model evidence. Journal of Computational Biology 19:418–438.

Arima, S. and L. Tardella. 2014. Inflated density ratio (IDR) method for estimating marginal likelihoods in Bayesian phylogenetics. chap. 3, Pages 25–57 *in* Bayesian phylogenetics: methods, algorithms, and applications (M.-H. Chen, L. Kuo, and P. O. Lewis, eds.). CRC Press, Boca Raton, Florida, USA.

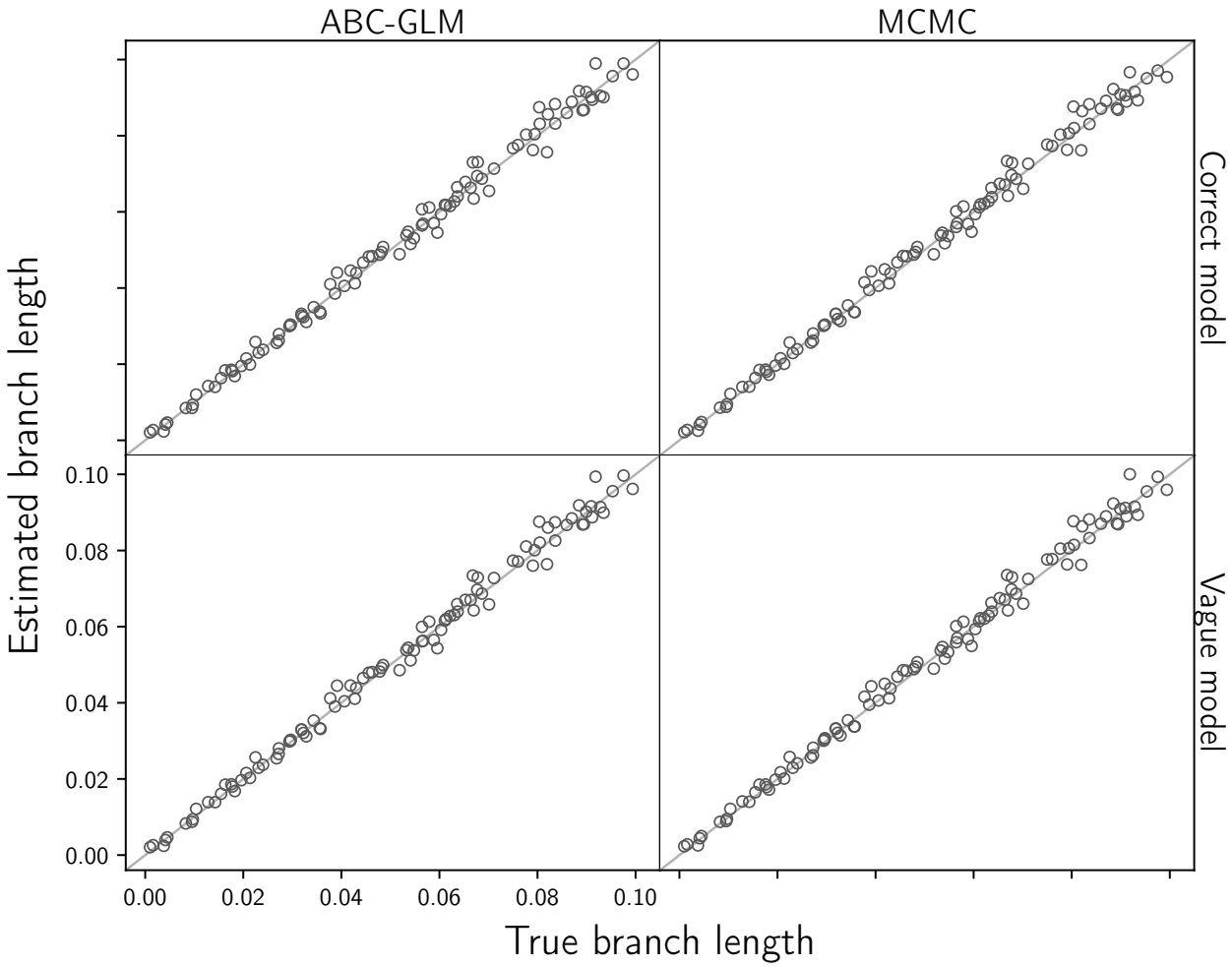

Figure A1. A comparison of the true branch length separating each pair of simulated sequences to the branch length estimated by ABC-GLM and full-likelihood MCMC under the correct branch-length model (branch length $\sim$ uniform(0.0001, 0.1)) and the vague model (branch length $\sim$ uniform(0.0001, 0.1)).

Baele, G. and P. Lemey. 2013. Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency. Bioinformatics 29:1970–1979.

Baele, G. and P. Lemey. 2014. Bayesian model selection in phylogenetics and genealogy-based population genetics. chap. 4, Pages 59–93 *in* Bayesian phylogenetics: methods, algorithms, and applications (M.-H. Chen, L. Kuo, and P. O. Lewis, eds.). CRC Press, Boca Raton, Florida, USA.

Baele, G., P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard, and A. V. Alekseyenko. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Molecular Biology and Evolution 29:2157–2167.

Baele, G., P. Lemey, and M. A. Suchard. 2016. Genealogical working distributions for Bayesian model testing with phylogenetic uncertainty. Systematic Biology 65:250–264.

Baele, G., P. Lemey, and S. Vansteelandt. 2013a. Make the most out of your samples: Bayes factor estimators for high-dimensional models of sequence evolution. BMC Bioinformatics 14:85.

Baele, G., W. L. S. Li, A. J. Drummond, M. A. Suchard, and P. Lemey. 2013b. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Molecular Biology and Evolution 30:239–243.

Beaumont, M., W. Zhang, and D. J. Balding. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025–2035.

Bemmels, J. B., P. O. Title, J. Ortego, and L. L. Knowles. 2016. Tests of species-specific models reveal the importance of drought in postglacial range shifts of a mediterranean-climate tree: insights from integrative distributional, demographic and coalescent modelling and ABC model selection. Molecular Ecology 25:4889–4906.

Berger, J. 2006. The case for objective Bayesian analysis. Bayesian Analysis 1:385–402.

Bergsten, J., A. N. Nilsson, and F. Ronquist. 2013. Bayesian tests of topology hypotheses with an example from diving beetles. Systematic Biology 62:660–673.

Bollback, J. P. 2002. Bayesian model adequacy and choice in phylogenetics. Molecular Biology and Evolution 19:1171–1180.

Bouchard-Côté, A. 2014. SMC (sequential Monte Carlo) for bayesian phylogenetics. chap. 8, Pages 163–185 *in* Bayesian phylogenetics: methods, algorithms, and applications (M.-H. Chen, L. Kuo, and P. O. Lewis, eds.). CRC Press, Boca Raton, Florida, USA.

Bouchard-Côté, A., S. Sankararaman, and M. I. Jordan. 2012. Phylogenetic inference via sequential Monte Carlo. Systematic Biology 61:579–93.

Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. PLOS Computational Biology 10:1–6.

Brandley, M. C., A. Schmitz, and T. W. Reeder. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. Systematic Biology 54:373–390.

Brown, J. M. and A. R. Lemmon. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Systematic Biology 56:643–655.

Brown, J. M. and R. C. Thomson. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Systematic Biology 66:517–530.

Carlin, B. P. and S. Chib. 1995. Bayesian model choice via Markov chain Monte Carlo methods. Journal of the Royal Statistical Society, Series B 57:473–484.

Chen, M.-H., Q.-M. Shao, and J. G. Ibrahim. 2000. Monte Carlo Methods in Bayesian Computation. Springer, New York, New York, USA.

Chiari, Y., V. Cahais, N. Galtier, and F. Delsuc. 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biology 10:65.

Del Moral, P. 1996. Non linear filtering: Interacting particle solution. Markov Processes and Related Fields 2:555–580.

Dinh, V., A. E. Darling, and F. A. Matsen IV. 2018. Online Bayesian phylogenetic inference: Theoretical foundations via Sequential Monte Carlo. Systematic Biology 67:503–517.

Fan, Y., R. Wu, M.-H. Chen, L. Kuo, and P. O. Lewis. 2011. Choosing among partition models in Bayesian phylogenetics. Molecular Biology And Evolution 28:523–532.

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

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

Fourment, M., B. C. Claywell, V. Dinh, C. McCoy, F. A. Matsen IV, and A. E. Darling. 2018. Effective online Bayesian phylogenetics via sequential Monte Carlo with guided proposals. Systematic Biology 67:490–502.

Geisser, S. 1980. In discussion of G. E. P. Box paper entitled: Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society, Series A 143:383–430.

Gelfand, A. E. and D. K. Dey. 1994. Bayesian model choice: asymptotic and exact calculations. Journal of the Royal Statistical Society Series B 56:501–514.

Gelfand, A. E., D. K. Dey, and H. Chang. 1992. Model determination using predictive distributions with implementation via sampling-based methods. Pages 147–167 *in* Bayesian

Statistics 4 (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds.). Oxford University Press, Oxford.

Gelfand, A. E. and S. K. Ghosh. 1998. Model choice: A minimum posterior predictive loss approach. Biometrika 85:1–11.

Gelman, A. and X.-L. Meng. 1998. Statistical Science 13:163–185.

George, E. I. and R. E. McCulloch. 1993. Variable selection via Gibbs sampling. Journal of the American Statistical Association 88:881–889.

Goldstein, M. 2006. Subjective Bayesian analysis: Principles and practice. Bayesian Analysis 1:403–420.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith. 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEE Proceedings F—Radar and Signal Processing 140:107–113.

Gray, R. R., A. J. Tatem, J. A. Johnson, A. V. Alekseyenko, O. G. Pybus, M. A. Suchard, and M. Salemi. 2011. Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant *Staphylococcus aureus* ST239 genome-wide data within a Bayesian framework. Molecular Biology and Evolution 28:1593–1603.

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

Grummer, J. A., R. W. Bryson Jr., and T. W. Reeder. 2013. Species delimitation using Bayes factors: simulations and application to the *Sceloporus scalaris* species group (Squamata: Phrynosomatidae). Systematic Biology 63:119–133.

He, Q., D. L. Edwards, and L. L. Knowles. 2013. Integrative testing of how environments from the past to the present shape genetic structure across landscapes. Evolution 67:3386–3402.

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

Holder, M. T., P. O. Lewis, D. L. Swofford, and D. Bryant. 2014. Variable tree topology stepping-stone marginal likelihood estimation. chap. 5, Pages 95–110 *in* Bayesian phylogenetics: methods, algorithms, and applications (M.-H. Chen, L. Kuo, and P. O. Lewis, eds.). CRC Press, Boca Raton, Florida, USA.

Hurvich, C. M. and C. Tsai. 1989. Regression and time series model selection in small samples. Biometrika 76:297–307.

Jeffreys, H. 1935. Some tests of significance, treated by the theory of probability. Proceedings of the Cambridge Philosophical Society 31:203–222.

Jeffreys, H. 1961. Theory of Probability. 3rd ed. Oxford University Press, Oxford, U.K.

Jones, G., Z. Aydin, and B. Oxelman. 2015. DISSECT: An assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. Bioinformatics 31:991–998.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. chap. 24, Pages 21–132 *in* Mammalian Protein Metabolism (H. N. Munro, ed.) vol. III. Academic Press, New York.

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

Knowles, L. L. and R. Massatti. 2017. Distributional shifts—not geographic isolation—as a probable driver of montane species divergence. Ecography .

Kuo, L. and B. Mallick. 1998. Variable selection for regression models. Sankhyā: The Indian Journal of Statistics, Series B 60:65–81.

Lad, F. 1996. Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction. John Wiley & Sons, Inc., New York, New York, USA.

Lartillot, N., H. Brinkmann, and H. Philippe. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evolutionary Biology 7:S4.

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

Leaché, A. D., M. K. Fujita, V. N. Minin, and R. R. Bouckaert. 2014. Species delimitation using genome-wide SNP data. Systematic Biology 63:534–542.

Leaché, A. D., M. T. McElroy, and A. Trinh. 2018a. A genomic evaluation of taxonomic trends through time in coast horned lizards (genus *Phrynosoma*). Molecular Ecology 27:2884–2895.

Leaché, A. D., T. Zhu, B. Rannala, and Z. Yang. 2018b. The spectre of too many species. Systematic Biology Page syy051.

Lepage, T., D. Bryant, H. Philippe, and N. Lartillot. 2007. A general comparison of relaxed molecular clock models. Molecular Biology And Evolution 24:2669–2680.

Leuenberger, C. and D. Wegmann. 2010. Bayesian computation and model selection without likelihoods. Genetics 184:243–252.

Lewis, P. O., W. Xie, M.-H. Chen, Y. Fan, and L. Kuo. 2014. Posterior predictive Bayesian phylogenetic model selection. Systematic Biology 63:309–321.

Lindley, D. V. 2000. The philosophy of statistics. The Statistician 49:293–337.

Liu, J. S. and R. Chen. 1998. Sequential Monte Carlo methods for dynamic systems. Journal of the American Statistical Association 93:1032–1044.

MacKay, D. J. C. 2005. Information Theory, Inference & Learning Algtorithms. 7.2 ed. Cambridge University Press, New York, New York, USA.

Massatti, R. and L. L. Knowles. 2016. Contrasting support for alternative models of genomic variation based on microhabitat preference: species-specific effects of climate change in alpine sedges. Molecular Ecology 25:3974–3986.

Maturana R., P., B. J. Brewer, S. Klaere, and R. R. Bouckaert. 2018. Model selection and parameter inference in phylogenetics using nested sampling. Systematic Biology Pages 1–20.

Mau, B. and M. A. Newton. 1997. Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. Journal of Computational and Graphical Statistics 6:122–131.

Minin, V. N., K. S. Dorman, F. Fang, and M. A. Suchard. 2005. Dual multiple change-point model leads to more accurate recombination detection. Bioinformatics 21:3034–3042.

Mueller, R. L., J. R. Macey, M. Jaekel, D. B. Wake, and J. L. Boore. 2004. Morphological homoplasy, life history evolution, and historical biogeography of plethodontid salamanders inferred from complete mitochondrial genomes. Proceedings of the National Academy of Sciences of the United States of America 101:13820–13825.

Neal, R. M. 2000. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics 9:249–265.

Newton, M. A. and A. E. Raftery. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. Journal of the Royal Statistical Society Series B-Statistical Methodology 56:3–48.

Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Systematic Biology 53:47–67.

Papadopoulou, A. and L. L. Knowles. 2016. Toward a paradigm shift in comparative phylogeography driven by trait-based hypotheses. Proceedings of the National Academy of Sciences 113:8018–8024.

Petris, G. and L. Tardella. 2007. New perspectives for estimating normalizing constants via posterior simulation. Tech. rep. Sapienza Università di Roma Roma, Italy.

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

Rannala, B. and Z. Yang. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. Systematic Biology 66:823–842.

Rubin, D. B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics 12:1151–1172.

Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics 6:461–464.

Skilling, J. 2006. Bayesian Analysis 1:833–859.

Suchard, M. A., C. M. R. Kitchen, J. S. Sinsheimer, and R. E. Weiss. 2003a. Hierarchical phylogenetic models for analyzing multipartite sequence data. Systematic Biology 52:649–664.

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

Suchard, M. A., R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer. 2003b. Inferring spatial phylogenetic variation along nucleotide sequences. Journal of the American Statistical Association 98:427–437.

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Molecular Biology And Evolution 18:1001–1013.

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2005. Models for estimating Bayes factors with applications to phylogeny and tests of monophyly. Biometrics 61:665–673.

Sugiura, N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. Communications in Statistics, Theory and Methods A7:13–26.

Sukumaran, J. and M. T. Holder. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly. 1997. Inferring coalescence times from DNA sequence data. Genetics 145:505–518.

Tolkoff, M. R., M. E. Alfaro, G. Baele, P. Lemey, and M. A. Suchard. 2018. Phylogenetic factor analysis. Systematic Biology 67:384–399.

Wang, L., S. Wang, and A. Bouchard-Côté. 2018a. An annealed sequential Monte Carlo method for Bayesian phylogenetics. arXiv:1806.08813 [q-bio.PE] .

Wang, Y.-B., M.-H. Chen, L. Kuo, and P. O. Lewis. 2018b. A new Monte Carlo method for estimating marginal likelihoods. Bayesian Analysis Pages 1–23.

Wegmann, D., C. Leuenberger, S. Neuenschwander, and L. Excoffier. 2010. ABCtoolbox: a versatile toolkit for approximate bayesian computations. BMC Bioinformatics 11:116.

Worobey, M., M. Gemmel, D. E. Teuwen, T. Haselkorn, K. Kunstman, M. Bunce, J.-J. Muyembe, J.-M. M. Kabongo, R. M. Kalengayi, E. Van Marck, M. T. P. Gilbert, and S. M. Wolinsky. 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455:661–664.

Wu, R., M.-H. Chen, L. Kuo, and P. O. Lewis. 2014. Consistency of marginal likelihood estimation when topology varies. chap. 6, Pages 113–127 *in* Bayesian phylogenetics: methods,

algorithms, and applications (M.-H. Chen, L. Kuo, and P. O. Lewis, eds.). CRC Press, Boca Raton, Florida, USA.

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

Figure Captions

Figure 1. An illustration of the posterior probability densities and marginal likelihoods of the four different prior assumptions we made in our coin-flipping experiment. The data are 50 "heads" out of 100 coin flips, and the parameter, θ, is the probability of the coin landing heads side up. The binomial likelihood density function is proportional to a Beta$(51, 51)$ and is the same across the four different beta priors on θ (M_1–M_4). The posterior of each model is a Beta$(\alpha+50, \beta+50)$ distribution. The marginal likelihoods ($P(D)$; the average of the likelihood density curve weighted by the prior) of the four models are compared.

Figure 2. A comparison of the approximate-likelihood Bayesian computation general linear model (ABC-GLM) estimator of the marginal likelihood (Leuenberger and Wegmann, 2010) to quadrature integration approximations (Xie et al., 2011) for 100 simulated datasets. We compared the ratio of the marginal likelihood (Bayes factor) comparing the correct branch-length model [branch length $\sim$ uniform(0.0001, 0.1)] to a model with a broader prior on the branch length [branch length $\sim$ uniform(0.0001, 0.2)]. The solid line represents perfect performance of the ABC-GLM estimator (i.e., matching the "true" value of the Bayes factor). The dashed line represents the expected Bayes factor when failing to penalize for the extra parameter space (branch length 0.1 to 0.2) with essentially zero likelihood. Quadrature integration with 1,000 and 10,000 steps using the rectangular and trapezoidal rule produced identical values of log marginal likelihoods to at least five decimal places for all 100 simulated datasets.

Figure A1. A comparison of the true branch length separating each pair of simulated sequences to the branch length estimated by ABC-GLM and full-likelihood MCMC under the correct branch-length model (branch length $\sim$ uniform(0.0001, 0.1)) and the vague model (branch length $\sim$ uniform(0.0001, 0.1)).