Title

Genealogical Working Distributions for Bayesian Model Testing with Phylogenetic Uncertainty

Permalink

https://escholarship.org/uc/item/5hf7q1z4

Journal

Systematic Biology, 65(2)

ISSN

1063-5157

Authors

Baele, Guy
Lemey, Philippe
Suchard, Marc A

Publication Date

2016-03-01

DOI

10.1093/sysbio/syv083

Peer reviewed

Syst. Biol. 65(2):250–264, 2016

DOI:10.1093/sysbio/syv083
Advance Access publication November 1, 2015

Genealogical Working Distributions for Bayesian Model Testing with Phylogenetic Uncertainty

Guy Baele[1,*], Philippe Lemey[1], and Marc A. Suchard[2,3,4]

[1]*Department of Microbiology and Immunology, Rega Institute, KU Leuven-University of Leuven, Leuven, Belgium;* [2]*Department of Biomathematics and* [3]*Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA;*
[4]*Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90095, USA*
Correspondence to be sent to: Department of Microbiology and Immunology, KU Leuven-University of Leuven, Minderbroedersstaat 10, 3000 Leuven, Belgium;
E-mail: guy.baele@rega.kuleuven.be.

Received 1 June 2015; reviews returned 27 October 2015; accepted 28 October 2015
Associate Editor: Mark Holder

Abstract.—Marginal likelihood estimates to compare models using Bayes factors frequently accompany Bayesian phylogenetic inference. Approaches to estimate marginal likelihoods have garnered increased attention over the past decade. In particular, the introduction of path sampling (PS) and stepping-stone sampling (SS) into Bayesian phylogenetics has tremendously improved the accuracy of model selection. These sampling techniques are now used to evaluate complex evolutionary and population genetic models on empirical data sets, but considerable computational demands hamper their widespread adoption. Further, when very diffuse, but proper priors are specified for model parameters, numerical issues complicate the exploration of the priors, a necessary step in marginal likelihood estimation using PS or SS. To avoid such instabilities, generalized SS (GSS) has recently been proposed, introducing the concept of "working distributions" to facilitate—or shorten—the integration process that underlies marginal likelihood estimation. However, the need to fix the tree topology currently limits GSS in a coalescent-based framework. Here, we extend GSS by relaxing the fixed underlying tree topology assumption. To this purpose, we introduce a "working" distribution on the space of genealogies, which enables estimating marginal likelihoods while accommodating phylogenetic uncertainty. We propose two different "working" distributions that help GSS to outperform PS and SS in terms of accuracy when comparing demographic and evolutionary models applied to synthetic data and real-world examples. Further, we show that the use of very diffuse priors can lead to a considerable overestimation in marginal likelihood when using PS and SS, while still retrieving the correct marginal likelihood using both GSS approaches. The methods used in this article are available in BEAST, a powerful user-friendly software package to perform Bayesian evolutionary analyses. [Bayes factor; Bayesian inference; coalescent model; marginal likelihood; MCMC; phylogenetics; Working distribution.]

The past decades have witnessed an increasing popularity of Bayesian inference in molecular phylogenetics, with a key role for Markov chain Monte Carlo (MCMC) in estimating posterior distributions under complex phylogenetic models (Yang and Rannala 1997). The computational demands associated with increasing model complexity and data quantity considered in modern phylogenetics have restricted the ability to assess model performance. In order to compare alternative models, a well-developed statistical theory such as model selection—which allows models to be evaluated according to objective criteria (Suchard et al. 2001; Huelsenbeck et al. 2001; Steel 2005)—should complement phylogenetic inference. Such approaches penalize the addition of extra parameters, unless there is a sufficiently impressive improvement in fit between model and data. The aim of model selection is hence not to find the true model that generated the data, but to select a model that best balances simplicity with flexibility and captures the key features of the biological process that generated the data (Steel 2005).

A standard approach to perform model selection in a Bayesian phylogenetic framework operates through the evaluation of Bayes factors (Sinsheimer et al. 1996; Suchard et al. 2001). The Bayes factor is a ratio of two marginal likelihoods (i.e., two normalizing constants of the form $p(Y|M)$, with Y the observed data and M an evolutionary model under evaluation), obtained for the two models, M_0 and M_1, under comparison (Jeffreys 1935):

$$B_{10} = \frac{p(Y|M_1)}{p(Y|M_0)}. \tag{1}$$

Although standard MCMC inference of posterior distributions avoids estimating the normalization constant or marginal likelihood $p(Y|M)$, it is of primary importance in evaluating model fit and calculating Bayes factors because it measures the average fit of a model to the data. Calculation of the marginal likelihood of model M requires integration of its likelihood across model parameter values Θ, weighted by the model's prior distribution

$$p(Y|M) = \int_{\theta\in\Theta} p(Y|\theta,M)\,p(\theta|M)\,\mathrm{d}\theta. \tag{2}$$

Among several models, one chooses the one of greatest marginal likelihood.

The introduction of path sampling (PS) into the fields of phylogenetics and molecular evolution has sparked renewed interest in estimating marginal likelihoods, which had been frequently approximated using a harmonic mean estimator (HME), but often with questionable results. Lartillot and Philippe (2006) compare PS to three variants of importance sampling

(IS): integrating the likelihood against the model prior (ILP), the HME and the stabilized HME. To this end, they use a Gaussian model with different dimensions and an evolutionary model on a fixed tree for which exact calculation of the marginal likelihood is available. In these comparisons, PS outperforms the IS variants across all scenarios, remaining well-behaved in cases with high dimensions where all three IS methods fail, even when using large numbers of costly posterior samples. Borrowing ideas from both IS and PS, Xie et al. (2011) further improve upon PS in their stepping-stone sampling (SS) approach and demonstrate that for a Gaussian model SS yields a substantially less biased estimator than PS. Importantly, SS also requires significantly fewer path steps than PS to estimate the marginal likelihood for realistic phylogenetic models with an acceptably small discretization bias.

Upon introduction, very little was known about the performance and computational issues of these methods, in particular when complex evolutionary and population genetic models needed to be fit to sizeable empirical data sets. To clarify these issues, Baele et al. (2012) specifically investigate the performance of PS and SS for comparing models of demographic change and relaxed molecular clocks based on both synthetic data and empirical examples. The authors show that PS and SS substantially outperform the posterior-based estimators (HME and sHME), leading PS and SS to correct erroneous conclusions drawn in previous analyses for three real-world data sets. Baele et al. (2012) also provide the implementation of these computationally demanding techniques into BEAST (Drummond et al. 2012), a cross-platform program for Bayesian analysis of molecular sequences via MCMC that offers a multitude of different models, such as autocorrelated and uncorrelated relaxed clock models, substitution models including heterogeneity across sites, coalescent models of population size and growth and phylogeographic models, with support for a flexible choice of prior specifications on model parameters. The availability of these techniques in a commonly used phylogenetic package has considerably contributed to a more widespread use in the field.

Despite these advances, the estimation of marginal likelihoods remains a challenging task, mostly because of computational restrictions. The BEAST implementation requires an initial burn-in to the posterior (for which the standard Bayesian analysis to estimate the models' parameters can be employed) and then evaluates a series of power posteriors along a path between posterior and prior using MCMC. In addition to the general computational burden of this routine, collecting samples from vague priors (or from distributions that are close, in the Kullback–Leibler (KL) sense, to the prior) through MCMC is notoriously difficult and time-consuming. Finally, numerical instabilities can arise when improper priors are used (Baele et al. 2013b), which lead to improper marginal likelihoods.

Recently, a new approach to estimate marginal likelihoods referred to as generalized stepping-stone sampling (GSS) has been proposed by Fan et al. (2011). This method generalizes the standard SS approach (Xie et al. 2011) by making use of a "working" distribution that is parameterized using samples from the posterior distribution. The authors show that if this working distribution exactly matches the posterior distribution, the marginal likelihood can be estimated exactly. GSS is considerably more efficient and does not require sampling from distributions close to the true prior, which can be problematic for vague prior specification. Despite the advantages GSS has to offer, it is currently of little use in a Bayesian coalescent-based framework that considers the genealogy to be unknown as it is currently restricted to fixed genealogies. Note, however, that Holder et al. (2014) provide a working distribution for tree topologies outside of a coalescent framework. Integrating over plausible genealogies complicates GSS because it requires defining a working distribution jointly for topologies and branch lengths that provides a good approximation to the posterior.

In this article, we propose two approaches to relax the restriction of fixing the underlying tree topology. A first approach constructs a matching "working" demographic distribution to the demographic model that is specified as a tree prior in a Bayesian genealogical analysis. The second approach aims at a more general design and constructs a product of exponential distributions based on the intercoalescent times of the underlying genealogy in each iteration. Using simulated Gaussian data, for which we can analytically calculate the true marginal likelihood, we show that GSS consistently estimates accurate marginal likelihoods even when employing very diffuse priors and outperforms the sHME, PS, and SS. For phylogenetic test cases where we are able to accurately estimate the true log marginal likelihood, we demonstrate similar superior performance of GSS. A large coalescent-based simulation study also reveals a higher accuracy for GSS compared with PS/SS when accommodating phylogenetic uncertainty, although not to the same extent as in the case of fixed topologies (Fan et al. 2011). Analyses of empirical data sets show that when assessing the model fit for standard demographic models and substitution models, PS/SS overestimate the marginal likelihood compared with GSS when very diffuse priors are used; that may influence the outcome of the model selection process.

Methods

Importance Sampling Estimators

Monte Carlo integration of the likelihood against the model prior (ILP), also known as independence sampling from the prior (ISP), produces an unbiased estimate of the (log) marginal likelihood. This importance sampling estimator uses the prior as importance sampling distribution:

$$p(Y|M) \simeq \frac{1}{K}\sum_{k=1}^{K} p(Y|\theta_k, M), \tag{3}$$

where $\{\theta_1,\ldots,\theta_K\}$ are independent draws from $p(\theta|M)$. This approach is rarely used when performing model selection in phylogenetics because it requires an enormous sampling effort to estimate marginal likelihoods, even when the data sets are limited. This is due to the observation that the high-likelihood region can be very concentrated. So, except for very large sampling efforts, the sample drawn from the prior is unlikely to contain sufficient draws from the high-likelihood region, resulting in a very poor estimate of the marginal likelihood (Lartillot and Philippe 2006).

The harmonic mean estimators (HME and sHME) constitute a class of (log) marginal likelihood estimators that only require samples from the posterior obtained by a standard Bayesian phylogenetic analyses using MCMC under a particular model (Newton and Raftery 1994). If one collects n samples $\{\theta_1,\ldots,\theta_n\}$ from the posterior, the HME is estimated as follows

$$p(Y|M) = \frac{n}{\sum_{k=1}^{n} \frac{1}{p(Y|\theta_k, M)}}. \tag{4}$$

To circumvent the HME's infinite variance in many practical situations, Newton and Raftery (1994) proposed the stabilized harmonic mean estimator (sHME), based on a mixture of the prior and the posterior, although in practice only samples from the posterior are used in computing the sHME.

PS Estimators

Most implementations of PS rely on drawing MCMC samples from a series of distributions, each of which is a power posterior differing only in its power, along the path going from the prior to the unnormalized posterior defined by the model M. Both Lartillot and Philippe (2006) and Xie et al. (2011) define this path to be:

$$q_\beta(\theta) = p(Y|\theta, M)^\beta p(\theta|M), \tag{5}$$

where $p(Y|\theta,M)$ is the likelihood function and $p(\theta|M)$ the prior. Hence, the power posterior is equivalent to the posterior distribution when $\beta = 1.0$ and reduces to the prior distribution when $\beta = 0.0$.

Different approaches have been proposed to determinine the values of β, that is the actual "powers" of the power posteriors from which one samples (Lartillot and Philippe 2006; Lepage et al. 2007; Friel and Petitt 2008). Xie et al. (2011) find that the efficiency of PS can be drastically improved by choosing β values according to evenly spaced quantiles of a Beta(α, 1.0) distribution rather than spacing β values evenly from 0.0 to 1.0; this represents a generalization of the approach by Friel and Petitt (2008). Xie et al. (2011) suggest using a value of $\alpha = 0.3$, which results in half of the β values evaluated being less than 0.1. The authors state that the positive skewness of this distribution is useful because (with sufficient and informative data) the likelihood only begins losing control over the power posterior for β values near 0, and at that point, the target distribution changes rapidly from something resembling the posterior to something resembling the prior. Xie et al. (2011) have shown that SS has better statistical properties and converges faster than PS, elevating it to the current model selection approach of choice in several software packages (Drummond et al. 2012; Ronquist et al. 2012).

Generalized stepping-stone sampling (GSS) involves constructing a path between the unnormalized posterior defined by the model M and a "working" distribution, that is in practice a product of independent probability densities parameterized using samples from the posterior distribution. Fan et al. (2011) define this path as:

$$q_\beta(\theta) = \left[p(Y|\theta, M)p(\theta|M)\right]^\beta \left[p_0(\theta|M)\right]^{1-\beta}, \tag{6}$$

where $p_0(\theta|M)$ is the working distribution. As with PS and SS, setting β to 1 yields the posterior, but setting β to 0 now yields the "working" distribution. Using a working distribution removes the problem of having to adequately sample from vague distributions (power posteriors near $\beta = 0$) in PS and SS. In addition, a working distribution that closely approximates the posterior yields a shorter path to integrate over and therefore involves less computational effort to accurately estimate the marginal likelihood (Fan et al. 2011).

Fan et al. (2011) propose an approach to match moments, for example, the marginal posterior sample mean and variance, to parameterize an independent working distribution for a parameter or block of parameters. For the nondemographic evolutionary parameters (see next section), we propose to use kernel density estimation (KDE) to compose the working distribution. KDE is a nonparametric approach to estimate the probability density function of a random variable. Let $X_1, X_2, \ldots, X_n$ denote a sample of size n of a random variable with unknown density f. The kernel density estimate of f at the point x is

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{7}$$

where the kernel K satisfies $\int K(x)dx = 1$ and the smoothing parameter h is known as the bandwidth (see e.g., Sheather (2004)). In practice, one generally chooses K to be a unimodal probability density on the same support as the original random variable after possible transformation. A popular choice for K is the normal kernel, namely:

$$K(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right). \tag{8}$$

We have implemented KDE in BEAST (Drummond et al. 2012) and consider normal kernels following appropriate transformation as necessary, with the bandwidth being automatically set at its optimal value h:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}}, \tag{9}$$

where $\hat{\sigma}$ is standard deviation of the transformed samples (Silverman 1986).

Matching Coalescent Model Working Distribution

Our first approach to construct a working distribution for the coalescent process focuses on providing a "working" coalescent model that summarizes the inferred parameters of the demographic prior assumption. We propose to first match the "working" coalescent model to the prior coalescent model, for example, when the prior coalescent model assumes a constant population size model, we also assume a constant population size model as the working distribution. Further, the parameters describing the working demographic distribution are set to their respective posterior empirical sample means, obtained from the parameter estimates of the preceding MCMC run, thereby informing the working distribution on plausible coalescent trees and hence shortening the path from posterior to working distribution in comparison to a diffuse prior. We denote this approach as generalized stepping-stone sampling using a matching coalescent model (GSS MCM). This approach readily applies to several parametric demographic models (in BEAST), but it may be cumbersome to match a flexible nonparametric coalescent prior—such as the Bayesian skyride model (Minin et al. 2008) or the Bayesian skygrid model (Gill et al. 2013)—due to its large number of parameters. To provide a working distribution for analyses incorporating more complicated demographic models, we have developed a second—more general—approach in the next section.

Product of Exponentials Working Distribution

Our second approach to construct a genealogical working distribution borrows ideas from the nonparametric Bayesian skyride model (Minin et al. 2008). We start with a genealogy $\mathbf{g}$ relating n sequences sampled at different time points $\mathbf{s}$, in units time. We describe the distribution for the genealogy with "heterochronous" sequence data as the general case, but it readily reduces to the simple case with contemporaneous tips. Coalescent theory provides a stochastic process that produces genealogies relating these sampled sequences. The process starts at sampling time $t=0$ and proceeds backward in time as t increases, coalescing n individuals one pair at a time until the time to the most recent common ancestor (TMRCA) of the sample is reached. Define the intercoalescent times $\mathbf{u}=(u_2,\dots,u_n)$ induced by $\mathbf{g}$, where $u_k=t_k-t_{k-1}$, t_k is the time of the $(n-k)$th coalescent event for $k=2,\dots,n$ and $t_n=0$ is the time of the most recently sampled sequence(s) (Fig. 1).

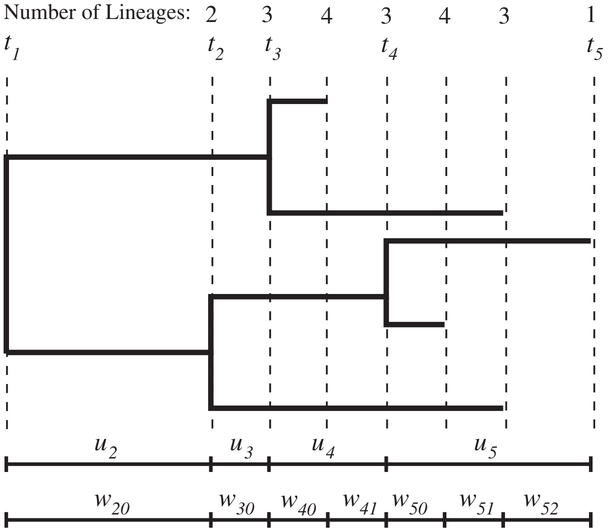

FIGURE 1. Example of a genealogy with intercoalescent interval notation. Times of coalescence and sampling events are depicted as vertical dashed lines with numbers of lineages present at these times shown above the lines. Below the genealogy, we mark the boundaries of intercoalescent intervals together with their lengths ($u_2,\dots,u_5$). We show how sampling events interrupt the intercoalescent intervals and produce subintervals with lengths ($w_{20},\dots,w_{52}$) at the bottom of the figure.

Branch lengths of $\mathbf{g}$ satisfy constraints imposed by the sampling times $\mathbf{s}$. The sampling times divide each intercoalescent interval k into subintervals $w_k=(w_{k0},\dots,w_{kj_k})$, where $j_k \in 0,\dots,n-1$ is the number of distinct sampling times occurring during interval k, with:

$$\sum_{j=0}^{j_k} w_{kj} = u_k, \tag{10}$$

where the interval that ends with the $(n-k)$th coalescent event always being indexed by $k0$. To each subinterval kj, we attach the number of lineages n_{kj} present in the genealogy at the beginning of this interval.

The number of intercoalescent times equals $n-1$ for n sampled sequences for any genealogy $\mathbf{g}$, regardless of the actual underlying bifurcating tree. We first collect m samples from the posterior for each intercoalescent time $u_k=t_k-t_{k-1}$ (with t_k in practice typically expressed in years). Let $\hat{\mu}_k$ be the posterior mean of the intercoalescent time k weighted by $\binom{n_k}{2}$ (with n_k the number of lineages in the k-th intercoalescent interval), and $\hat{\mu}=(\hat{\mu}_2,\dots,\hat{\mu}_n)$:

$$\hat{\mu}_k = \frac{1}{m}\sum_{k=1}^{m}\binom{n_k}{2}(t_k-t_{k-1}). \tag{11}$$

In the event that one or more sampling events occur during intercoalescent time k, the expression for $\hat{\mu}_k$

becomes:

$$\hat{\mu}_k = \frac{1}{m}\sum_{k=1}^{m}\frac{\sum_{j=0}^{j_k} n_{kj}(n_{kj}-1)w_{kj}}{2}. \tag{12}$$

In other words, in each iteration of the posterior exploration we keep track of the maximum-likelihood estimate (MLE) of each of the effective population sizes $\hat{\theta}_k$ of the Bayesian skyride model (see Equation (5) in Minin et al. (2008)):

$$\hat{\theta}_k = \frac{\sum_{j=0}^{j_k} n_{kj}(n_{kj}-1)w_{kj}}{2}. \tag{13}$$

The estimates of $\hat{\theta}_k$ are then smoothed using a LOESS (local polynomial regression fitting) estimator, to mimic the smoothing prior used in the Bayesian skyride model. The resulting values are subsequently applied as $\hat{\mu}_k$ in Equation 17 (see below) to construct the working distribution.

Let $\hat{\phi}_k$ be the empirical variance of the intercoalescent time k weighted by $\binom{n_k}{2}$ (with n_k the number of lineages in the k-th intercoalescent interval), and $\hat{\phi} = (\hat{\phi}_2, \ldots, \hat{\phi}_n)$:

$$\hat{\phi}_k = \frac{1}{m}\sum_{k=1}^{m}\left[\binom{n_k}{2}(t_k - t_{k-1}) - \hat{\mu}_k\right]^2. \tag{14}$$

As in Minin et al. (2008), we distinguish between coalescent and sampling events. In our notation, subintervals labeled as $k0$ end with a coalescent event. Each such subinterval contributes an exponential density to the coalescent likelihood, where the exponential rate depends on the number of lineages present and the empirical posterior sample mean collected for that interval. Because in our notation only subintervals with indices $k0$ end with a coalescence event, this contribution equals:

$$\frac{n_{k0}(n_{k0}-1)}{2\hat{\mu}_k}\exp\left[-\frac{n_{k0}(n_{k0}-1)w_{k0}}{2\hat{\mu}_k}\right]. \tag{15}$$

Subintervals ending with a sampling event contribute a probability of no coalescence to the likelihood, or equivalently, the probability that an exponentially distributed coalescence time is greater than the interval length, that is for each such subinterval with index kj:

$$\exp\left[-\frac{n_{kj}(n_{kj}-1)w_{kj}}{2\hat{\mu}_k}\right]. \tag{16}$$

Hence, the likelihood of observing subintervals w_k comprising intercoalescent interval k is

$$\Pr(w_k \,|\, \hat{\mu}_k) = \frac{1}{\hat{\mu}_k}\exp\left[-\frac{\sum_{j=0}^{j_k} n_{kj}(n_{kj}-1)w_{kj}}{2\hat{\mu}_k}\right], \tag{17}$$

where the first binomial was dropped because we consider the tree topology as random (Rodrigo and Felsenstein 1999), and with the values for $\hat{\mu}_k$ obtained from the LOESS estimator.

By taking the log:

$$\log\left[\Pr(w_k \,|\, \hat{\mu}_k)\right] = -\log(\hat{\mu}_k) - \sum_{j=0}^{j_k}\frac{n_{kj}(n_{kj}-1)w_{kj}}{2\hat{\mu}_k}, \tag{18}$$

and summing over k we arrive at the following log density:

$$\log\left[\Pr(w \,|\, \hat{\mu})\right] = \sum_{k=2}^{n}\log\left[\Pr(w_k \,|\, \hat{\mu}_k)\right]. \tag{19}$$

This (coalescent) density will serve as the working demographic distribution for the coalescent process. We denote this approach as generalized stepping-stone sampling using a product of exponentials with LOESS smoothing (GSS POEL).

Examples

Simulated Gaussian Example

We first compare different marginal likelihood estimators on a simple Gaussian example for which closed-form expressions of the marginal likelihoods are available. This allows for an objective comparison between the various methods. We perform a simulation experiment to compare the performance of the ILP, sHME, PS, SS, and the proposed GSS method. Suppose n observations are sampled from a normal distribution with mean μ and precision τ. Let $Y = (y_1, \ldots, y_n)$ be the data. The likelihood can be written in the following form:

$$p(Y \,|\, \mu, \tau) = \frac{1}{(2\pi)^{n/2}}\tau^{n/2}\exp\left(-\frac{\tau}{2}\left[n(\mu-\bar{x})^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2\right]\right). \tag{20}$$

The conjugate prior is the Normal-Gamma:

$$NG(\mu, \tau \,|\, \mu_0, \kappa_0, \alpha_0, \beta_0) \overset{\text{def}}{=} N(\mu \,|\, \mu_0, (\kappa_0\tau)^{-1})G(\tau \,|\, \alpha_0, \beta_0). \tag{21}$$

As shown in Murphy (2007), the posterior equals

$$\begin{aligned} p(\mu, \tau \,|\, Y) &= NG(\mu, \tau \,|\, \mu_n, \kappa_n, \alpha_n, \beta_n), \text{ with} \\ \mu_n &= \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0 + n} \\ \kappa_n &= \kappa_0 + n \\ \alpha_n &= \alpha_0 + n/2 \\ \beta_n &= \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\kappa_0 n(\bar{x} - \mu_0)^2}{2(\kappa_0 + n)}, \end{aligned} \tag{22}$$

and the closed-form expression for the marginal likelihood becomes:

$$p(Y) = \frac{\Gamma(\alpha_n)}{\Gamma(\alpha_0)}\frac{\beta_0^{\alpha_0}}{\beta_n^{\alpha_n}}\left(\frac{\kappa_0}{\kappa_n}\right)^{\frac{1}{2}}(2\pi)^{-n/2}. \tag{23}$$

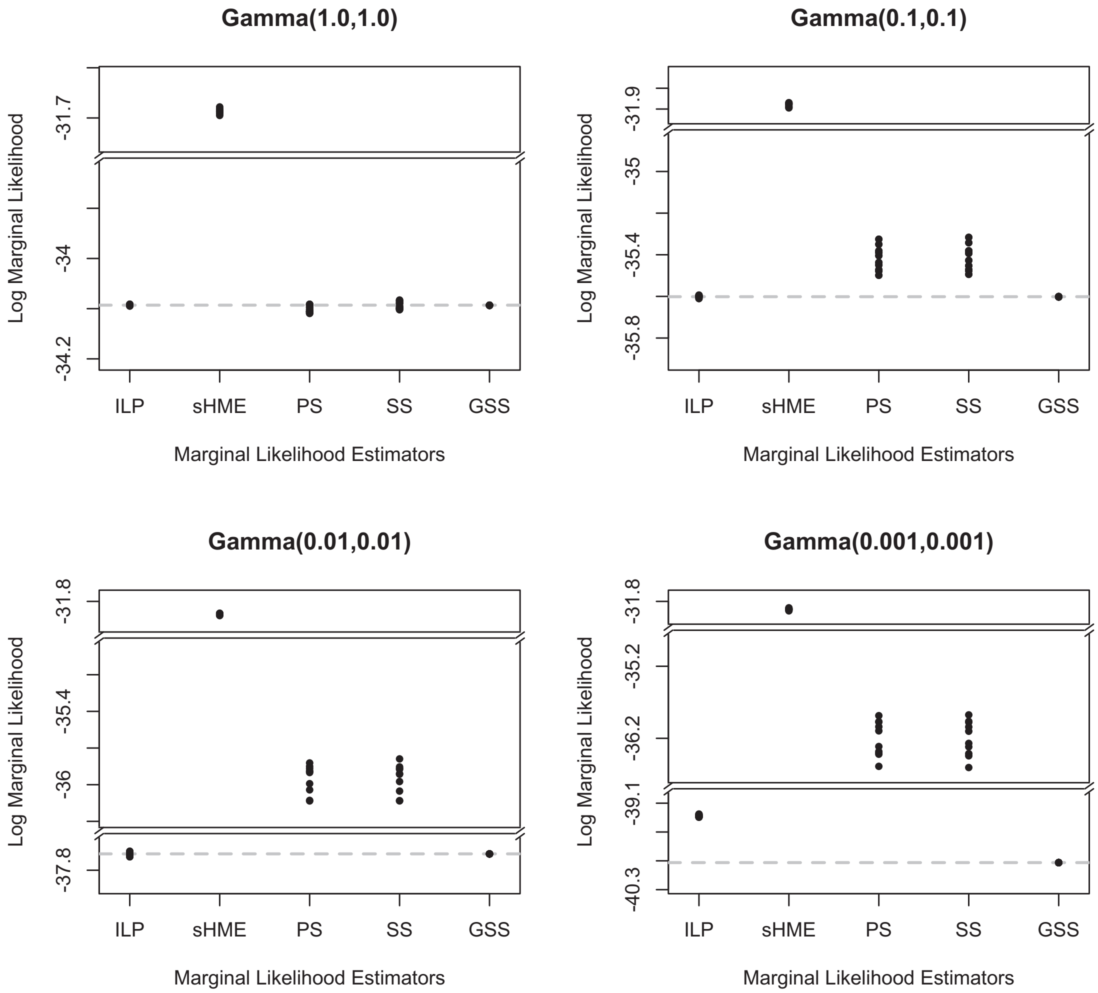

FIGURE 2. Log marginal likelihood estimates for Gaussian examples simulated under a normal distribution with four different gamma priors on the precision. For these data sets, the true (log) marginal likelihood can be calculated analytically (Murphy 2007). This value is indicated by a dashed gray line for each of the gamma priors tested. Five different estimators were used, with a number of different settings for each estimator: integrating the likelihood against the prior (ILP), the smoothed harmonic mean estimator (sHME), path sampling (PS), stepping-stone sampling (SS), and generalized stepping-stone sampling (GSS). This example shows that GSS is by far the most accurate approach to estimate (log) marginal likelihood for all the gamma priors tested. The sHME systematically overestimates the true (log) marginal likelihood, whereas PS and SS do so as the gamma prior becomes increasingly uninformative, but not to the same extent as for the sHME.

We simulate a single data set of size $n = 20$ from a normal distribution having mean $\mu = 0.0$ and precision $\tau = 1.0$. The prior for μ is normally distributed with mean $\mu_0 = 2.0$ and precision τ, which in turn is equipped with a gamma-distributed prior with shape parameter α_0 and rate parameter β_0. Here, we are particularly interested in the influence of the precision prior on the ability of the various marginal likelihood estimators to retrieve the true value. In Bayesian phylogenetic inference, vague or uninformative but proper priors are often used on most parameters due to the lack of available prior information (Baele et al. 2013b). To test which priors complicate the estimation of the marginal likelihood, we start with a relatively diffuse gamma prior and further decrease its informativeness in a gradual way. In particular, we test the following priors for τ: $G(\alpha_0 = \beta_0 = 1.0)$, $G(\alpha_0 = \beta_0 = 0.1)$, $G(\alpha_0 = \beta_0 = 0.01)$ and $G(\alpha_0 = \beta_0 = 0.001)$.

We performed the analyses on the simulated Gaussian data using the ILP, sHME, PS, SS, and GSS marginal likelihood estimators (Fig. 2). We rely on MCMC approximation for all the estimators, except for the ILP, which draws new parameter values directly from the prior. We run the ILP and sHME for 25 million iterations, whereas the PS, SS, and GSS runs explore 25 power posteriors, each with an MCMC run of 1 million iterations. We define the path of power posteriors p_β from posterior to (working) prior for this latter set of estimators according to evenly spaced quantiles of a Beta(α, 1.0) distribution, with $\alpha = 0.3$, as suggested by Xie et al. (2011).

In agreement with previous studies (Lartillot and Philippe 2006; Xie et al. 2011; Baele et al. 2012, 2013b; Baele and Lemey 2013), we find that the sHME systematically overestimates the marginal likelihood, independent of the prior choice on τ. For the least diffuse gamma prior $G(1.0,1.0)$, all other estimators are able to retrieve the true value of the marginal likelihood, albeit with varying accuracy. GSS stands out as the estimator that consistently (and most accurately) yields accurate marginal likelihood estimates for all prior choices. As the gamma prior becomes more diffuse, the PS and SS estimators overestimate the true marginal likelihood by a larger margin. For the most diffuse gamma prior $G(0.001,0.001)$, a popular choice as vague prior in the Bayesian phylogenetics community, the GSS estimator still succeeds in retrieving the true value, with remarkable accuracy. The ILP performs well in all but one scenario, that is when the gamma prior $G(0.001,0.001)$ is assumed, which is notoriously difficult to sample from. We note that the ILP is able to draw samples directly from the prior rather than approximating the prior using MCMC. PS and SS on the other hand resort to MCMC to approximate the power posteriors all the way down to the prior, and poor MCMC sampling from diffuse priors may result in less accurate marginal likelihood estimates for these procedures.

Phylogenetic Examples

Whereas a Gaussian test case allows us to analytically calculate the true value of the (log) marginal likelihood, it offers little predictive power as to how these estimators will perform in a Bayesian phylogenetic framework. We therefore first explore small phylogenetic test cases, consisting of four data sets with respectively 3, 4, 5, and 6 taxa, drawn at random from the intergenic *Staphylococcus aureus* data set of Gray et al. (2011). We first performed a standard Bayesian inference through MCMC using BEAST (Drummond et al. 2012) to estimate the parameters of a constant population size model, an HKY substitution model (Hasegawa et al. 1985) and a strict clock model, while estimating the tree topology and branch lengths. Because our main interest lies in accommodating phylogenetic uncertainty in the GSS estimation procedure, we fix most of these parameters to their mean posterior value, except the HKY's transition/transversion ratio parameter, the tree topology, and branch lengths, which are allowed to vary in the different marginal likelihood estimations. We provide a simpler test case in Supplementary Material available on Dryad (http://dx.doi.org/10.5061/dryad.8tm76).

In order to determine which of the approaches yields adequate performance in a phylogenetic setting, we require the true value of the (log) marginal likelihood for each of the four data sets. To this end, we use the ILP estimator and repeatedly sample from the coalescent prior, that is a constant population size model with a fixed population size, and from the prior on the transition/transversion ratio (a relatively diffuse Gamma(0.01, 0.01) prior). This approach yields a marginal likelihood value that converges onto a specific value for the proposed phylogenetic data sets (Supplementary Figure S1 available on Dryad (http://dx.doi.org/10.5061/dryad.8tm76), showing that the true value of the marginal likelihood can be approximated relatively well, albeit at a very high computational cost that increases with the size of the data set and the complexity of the model.

We proceed by testing various (log) marginal likelihood estimators on each of these data sets. For the HME and sHME, we run a total of 25 million MCMC iterations, using the default transition kernels on the tree topology and branch lengths. For the PS, SS, GSS MCM, and GSS POEL estimators, we assume a path from posterior to (working) prior that consists of 25 power posteriors, distributed according to evenly spaced quantiles of a Beta(0.3, 1.0) distribution, and for each power posterior we run an MCMC analysis of 1 million iterations. For the GSS estimators, a working distribution on the transition/transversion parameter is constructed using the KDE approach with normal kernels (after appropriate transformations), from the samples collected during its posterior exploration. We summarize the performance of the different marginal likelihood estimators, reporting their mean, standard deviation (SD), and root mean square error (RMSE) for 25 independent runs (Table 1). The RMSE is defined as

$$\mathrm{RMSE}=\sqrt{\mathbb{E}(\log\hat{r}-\log r_{\mathrm{true}})^2}.$$

Both the HME and sHME systematically overestimate the estimated log marginal likelihood to a large extent, leading to high RMSE values (Table 1). This overestimation is more pronounced as the number of taxa increases. The path sampling class of estimators (i.e., PS and SS) result in much smaller RMSE values than the HME and sHME, but still fail to retrieve the true log marginal likelihood for all four data sets; PS and SS offer highly comparable performance, but both overestimate the (log) marginal likelihood to some extent. The GSS estimators clearly outperform PS/SS, being better able to retrieve the true value than the latter. Both GSS estimators yield similar performance, as indicated by the reported RMSE values. The overestimation by PS/SS can be reduced by drastically increasing the computational settings (i.e., the number of power posteriors and the chain length per power posterior), as is shown in Supplementary Material available on Dryad (http://dx.doi.org/10.5061/dryad.8tm76). However, even increasing these settings 100-fold still does not yield similar performance as the class of GSS estimators.

In the next section, we provide a more thorough investigation of the performance of the various (log) marginal likelihood estimators using larger simulated data sets while integrating out a typical set of evolutionary parameters, as well as accommodating phylogenetic uncertainty.

Simulated Phylogenetic Data

Following our previous work (Baele et al. 2012), we also simulate phylogenetic data in order to assess the

TABLE 1. Small phylogenetic test examples, containing 3, 4, 5, and 6 sequences from a previously published data set

3 Taxa; True value: −1895.095				4 Taxa; True value: −1938.851			
	K = 25; C == 1M				K = 25; C = 1M		
Method	Mean	SD	RMSE	Method	Mean	SD	RMSE
HME	−1887.286	0.663	7.836	HME	−1929.881	0.880	9.011
sHME	−1886.203	0.020	8.892	sHME	−1928.409	0.025	10.441
PS	−1894.543	0.559	0.778	PS	−1938.344	0.536	0.730
SS	−1894.625	0.670	0.807	SS	−1938.371	0.638	0.788
GSS MCM	−1895.097	0.029	0.029	GSS MCM	−1938.852	0.031	0.030
GSS POEL	−1895.099	0.020	0.020	GSS POEL	−1938.855	0.022	0.021
5 Taxa; True value: −2129.607				6 Taxa; True value: −2293.076			
	K = 25; C = 1M				K = 25; C = 1M		
Method	Mean	SD	RMSE	Method	Mean	SD	RMSE
HME	−2120.108	0.969	9.546	HME	−2279.906	0.419	13.176
sHME	−2118.262	0.031	11.346	sHME	−2278.225	0.029	14.851
PS	−2129.286	0.611	0.679	PS	−2293.030	0.734	0.721
SS	−2129.195	0.646	0.755	SS	−2292.923	0.827	0.825
GSS MCM	−2129.612	0.035	0.035	GSS MCM	−2293.079	0.031	0.032
GSS POEL	−2129.597	0.031	0.033	GSS POEL	−2293.072	0.040	0.042

Except for the transition/transversion rate ratio κ of the HKY model, for which we specify a Gamma(0.01, 0.01) prior, all other parameters (constant population size and strict clock rate) are set to their mean value obtained from an initial MCMC run. The coalescent tree and κ are being sampled/updated during the runs in this table. Mean, SD, and RMSE from 25 independent runs in BEAST are shown throughout the table. K indicates the number of power posteriors for PS/SS/GSS, with C the chain length per power posterior. The HME and sHME systematically overestimate the true log marginal likelihood to a large extent. PS and SS now also systematically overestimate the true marginal likelihood, which can be attributed to the diffuse prior on κ. Only the GSS estimators are able to accurately retrieve the true value of the log marginal likelihood.

operating characteristics of the different (log) marginal likelihood estimators. Based on the coalescent analysis of Worobey et al. (2008), we consider the sampling dates of 60 sequences that represent the diversity in the original HIV-1 group M data set and simulate dated-tip genealogies under a simple constant population size model. We simulate 100 genealogies under this scenario using CoalGen, which is part of the BEAST software package (Drummond et al. 2012). Along each genealogy, we simulate sequences encompassing 1000 sites using GTR parameter values (Tavaré 1986), varying rates across sites—modelled using a discretised gamma distribution (Yang 1996)—and a substitution rate that reflects the estimates for the real data (Bielejec et al. 2014).

For each simulated data set under each demographic model, we employ seven different approaches to estimate the log marginal likelihood: the HME, sHME, PS, SS, GSS Fixed (with a fixed tree topology and therefore not requiring a working distribution for this parameter), GSS MCM (with a random tree topology and a constant population size model as its working demographic distribution), and GSS POEL (a product of exponentials with LOESS smoothing as working distribution). For all marginal likelihood estimators, we ran the same amount of 5×10^7 MCMC iterations in BEAST to ensure a fair comparison (not including initial burn-in to the posterior nor collection of the samples required to construct the working distributions). For the HME and sHME, this means running a standard Bayesian inference by using MCMC for 50 million iterations, whereas for all other estimators, 50 power posteriors were run for 1 million iterations each, along a path defined by a Beta$(0.3, 1.0)$ distribution. We run each estimation procedure twice, with different starting values for the models' parameters, in order to test the repeatability of the various methods (Figs. 3 and 4).

We first test the repeatability of the HME, sHME, and GSS Fixed (Fig. 3). To perform an objective comparison, we propose a simple summary statistic: the average over all 100 simulated data sets of the absolute difference in log marginal likelihood for two independent estimates:

$$D=\mathbb{E}\left(|\,\mathrm{MLE}_1-\mathrm{MLE}_2\,|\right). \tag{24}$$

The HME has in theory an infinite variance, explaining why it suffers from poor repeatability, with differences in log marginal likelihood between two independent runs as high as nearly 10 log units. The stabilized or smoothed HME remedies this problem and allows for higher repeatability among independent runs. This also holds true for the GSS Fixed estimator, which operates

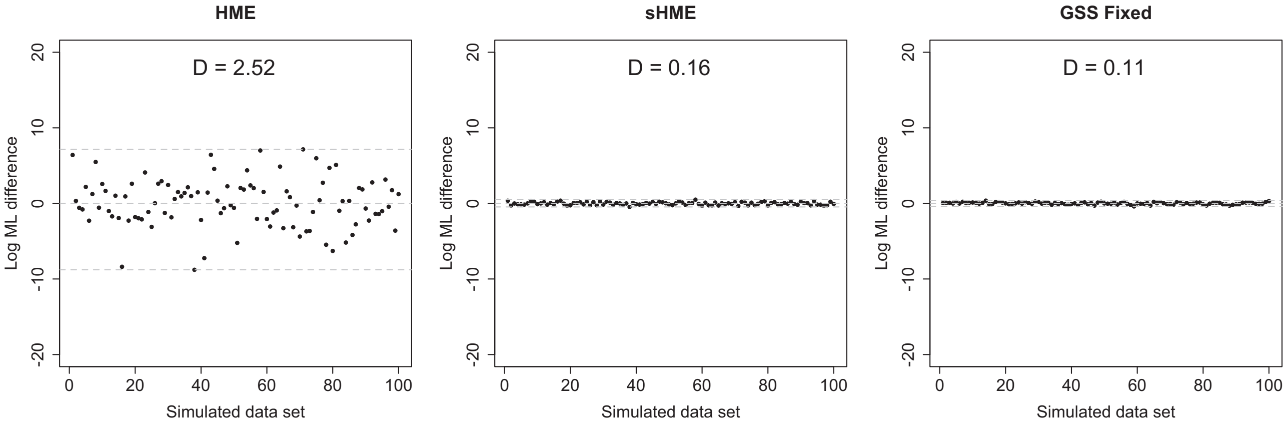

FIGURE 3. Repeatability plots for the harmonic mean estimator (HME), stabilized HME (sHME) and generalized stepping-stone sampling assuming a fixed tree topology (GSS Fixed), based on 100 simulated data sets and two independent runs employing different starting values. The repeatability of the HME is considerably lower than that of the sHME and GSS Fixed. One should be cautious concerning the high repeatability of the sHME—as it systematically overestimates the log marginal likelihood—and GSS Fixed—as it does not accommodate phylogenetic uncertainty and as a consequence provides a different estimate of the log marginal likelihood, knowing the tree under which the data were simulated.

under the restrictive assumption of a fixed tree topology, rendering comparisons unfair. Given that the HME and sHME systematically overestimate the log marginal likelihood and the GSS Fixed estimator provides a different estimate of the log marginal likelihood because it does not incorporate phylogenetic uncertainty and is hence performed on the tree that was used to simulate the data, we focus on the performance of the PS, SS, GSS MCM, and GSS POEL estimators that do accommodate phylogenetic uncertainty.

Both PS and SS are clearly outperformed by the GSS estimators in terms of repeatability/variance between runs (Fig. 4). Counterintuitively, PS seems to have better repeatability than SS. However, PS is clearly biased compared with SS, due to its discretization bias (Xie et al. 2011). Both GSS MCM and GSS POEL outperform PS and SS in terms of repeatability, using identical computational settings, while accommodating phylogenetic uncertainty. The repeatability statistic indicates a slight advantage for the GSS POEL estimator. Compared with the GSS POEL approach, the HME and sHME approaches appear to overestimate the log marginal likelihood when accommodating phylogenetic uncertainty (see Supplementary Materials available on Dryad http://dx.doi.org/10.5061/dryad.8tm76). Not accommodating phylogenetic uncertainty when using the GSS Fixed approach also yields higher log marginal likelihoods, but this does not necessarily represent an overestimation.

Empirical Examples

The epidemic history of HIV-1.—To assess the performance of the new marginal likelihood estimators on empirical examples, we revisit a Bayesian evolutionary reconstruction of the HIV-1 group M epidemic history originally performed by Worobey et al. (2008). This study examines sequence data from a 1960 specimen from Leopoldville in the Belgian Congo (now Kinshasa, Democratic Republic of the Congo) that shows considerable divergence from the 1959 (ZR59) sequence (Zhu et al. 1998), the oldest and only known sequence sampled before 1976 at that time. The authors show that the inclusion of the 1959 and 1960 sequences appears to improve estimation of the TMRCA of the M group, limiting the influence of the coalescent tree prior on the posterior TMRCA distributions. However, scientific interest also lies in characterizing the HIV-1 group M population dynamics through time as captured by different coalescent models.

Worobey et al. (2008) consider different coalescent models, both parametric and nonparametric, as prior distributions for time-measured trees. Our previous work has shown that, for this data set, the constant population size model fits the data significantly worse than the other coalescent models considered, but a consistent difference in performance between the other coalescent models could not be established, even with considerable computational investment (Baele and Lemey 2014). We revisit this HIV-1 data set using four coalescent models: the constant population size model, the exponential growth model, the expansion growth model, and a recently developed two-phase exponential-logistic growth model (Faria et al. 2014). The latter model estimates growth rate parameters for each growth period independently and provides an estimate of the time of transition between the exponential and logistic periods. We estimate log marginal likelihoods for these models using PS, SS, GSS MCM, and GSS POEL (Fig. 5). In these analyses, we fix the number of path steps to 64 and gradually increase the chain length per path step until convergence has been reached.

PS consistently appears to overestimate marginal likelihoods as compared with SS when using identical

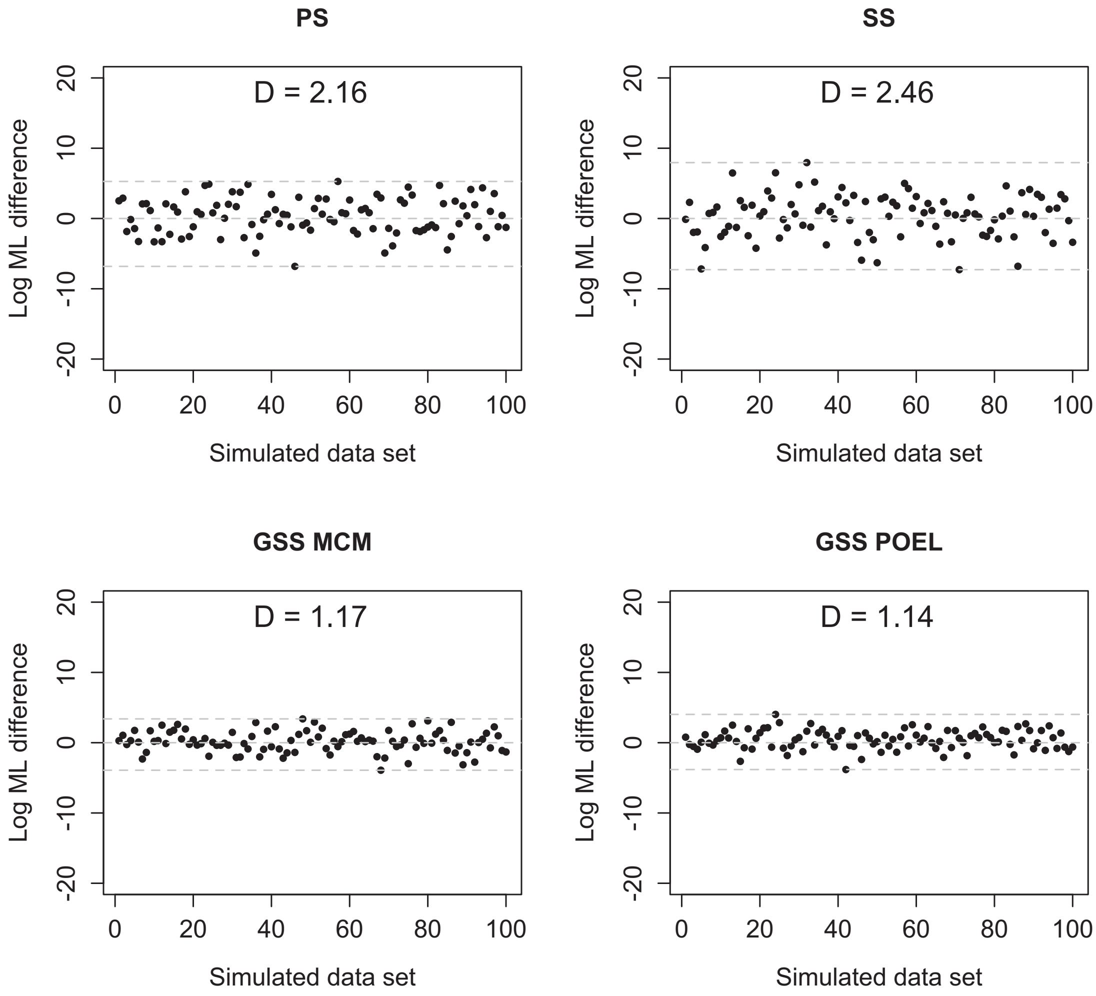

FIGURE 4. Repeatability plots for PS, SS, generalized stepping-stone sampling using a constant population size model as working distribution (GSS MCM) and generalized stepping-stone sampling using a product of exponentials with LOESS smoothing as working distribution (GSS POEL). The difference between two independent runs, employing different starting values, across 100 simulated data sets are shown. This suggests that the previously published low variance for GSS is mainly due to fixing the tree topology. When relaxing this assumption, however, GSS still has lower variance between runs than PS and SS, indicating its increased accuracy over those methods. Both GSS implementations have similar repeatability.

computational settings (Fig. 5), in line with previous conclusions by Xie et al. (2011). The overestimation is relatively constant, between 3 and 5 log units for each demographic model, hence not affecting the outcome of the model selection, when compared with SS. In turn, SS (and by extension PS) seem to consistently overestimate the log marginal likelihood when contrasted against the GSS MCM and GSS POEL estimates. This overestimation is, however, not constant and can affect the outcome of comparison of demographic models. SS, GSS MCM, and GSS POEL consider the two single-phase growth models, that is exponential and expansion, to be quite similar in terms of model fit, as their log marginal likelihoods only vary between 0.5 and 3.8 log units (with the constant population size model performing far worse, yielding a difference of around 160 log units). Both GSS MCM and GSS POEL consider the exponential-logistic growth model to perform significantly better in terms of model fit (BF >10) than the exponential and expansion growth models. This is in line with the epidemic history of HIV-1 group M, as reconstructed using a nonparametric demographic model, which has periods of exponential and logistic growth (Worobey et al. 2008). Even when assuming proper priors on all the parameters of the exponential-logistic growth model (Faria et al. 2014), PS and SS fail to complete their exploration of the power posteriors close to the prior—due to numerical integration problems associated with the demographic function—and hence fail to provide a log marginal likelihood for this model. This points to another advantage of our proposed GSS MCM and GSS POEL approaches: they avoid the exploration of such vague distributions altogether. Comparing these four demographic models using GSS reveals that the exponential-logistic growth model outperforms the two other growth models by about 10 log units, with the constant population size model yielding a much lower fit than the other models.

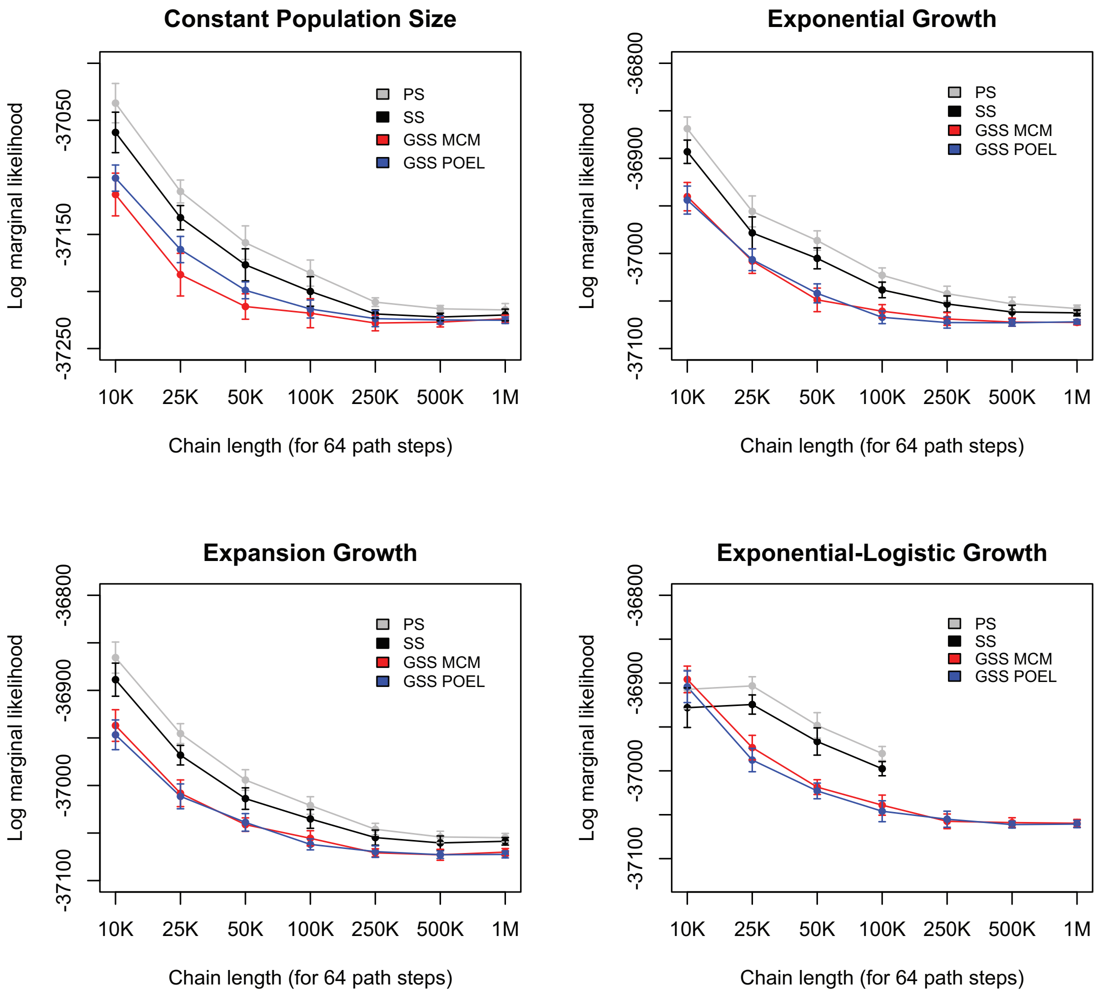

FIGURE 5. Convergence assessment of PS, SS, GSS MCM, and GSS POEL estimators on the HIV-1 data example of Worobey et al. (2008). A fixed number of 65 power posteriors, required to construct 64 path steps, were run along the path between posterior and (working) prior for all (log) marginal likelihood estimators, assuming different chain lengths per power posterior. Ten replicates were run for each computational setting and for each demographic model. The mean of these replicates is plotted along with the standard deviation. PS and SS consistently overestimate the log marginal likelihood when contrasted against GSS MCM and GSS POEL. In general, both GSS methods converge faster, with less iterations per power posterior, to a stable log marginal likelihood. Estimating the log marginal likelihood of the exponential-logistic growth model fails using PS/SS for more demanding computational settings, even with proper priors on all its parameters. For the most demanding computational settings (but also for most of the other settings), the GSS approach that employs a product of exponentials with LOESS smoothing (GSS POEL) has lower variance than the GSS approach that matches the demographic model as its working distribution (GSS MCM).

Both GSS MCM and GSS POEL offer increased precision compared with SS (and also PS, which we do not discuss because SS converges faster), with GSS POEL consistently outperforming the GSS MCM approach. This increase in precision is about 3% for the expansion growth model, seemingly in line with the repeatability findings of our phylogenetic simulation study, but reaches higher levels for the other demographic models: 13% for the exponential-logistic growth model, 29% for the constant population size model, and 35% for the exponential growth model. We therefore conclude that for an empirical example, GSS POEL also emerges as the preferred (log) marginal likelihood estimator. Comparing the precision of this approach to SS, we observe an increase of 14% for the expansion growth model, 78% for the exponential growth model, and 214% for the constant population size model (with again no basis of comparison for the exponential-logistic growth model). These statistics show the increase in accuracy of our proposed GSS approaches compared with existing state-of-the-art (log) marginal likelihood estimators, such as PS/SS.

Finally, we present timing assessments for the different marginal likelihood estimators that were used to analyze the HIV-1 group M data set (Fig. 5). Such a comparison of run times illustrates that GSS approaches require less computation time compared with PS/SS (Fig. 6). This may seem counterintuitive because the GSS estimation process requires the evaluation of a potentially large amount of working distributions. However, the computational advantage of both GSS approaches can be attributed to the absence of the numerical instabilities found in PS/SS. Whereas PS/SS often encounter loss of precision when

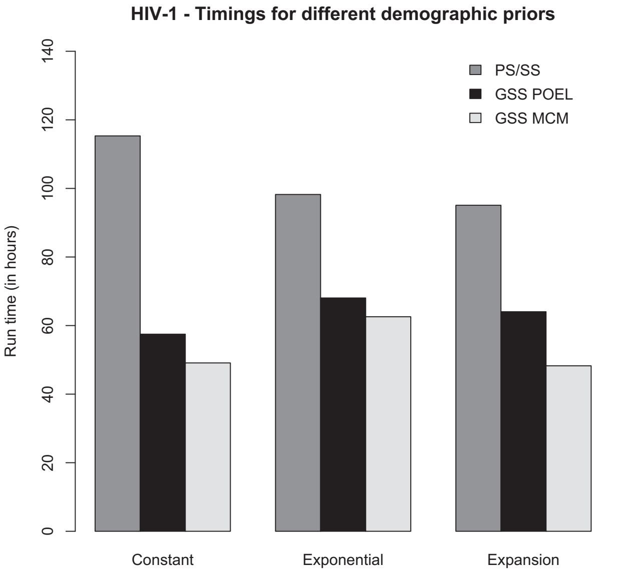

FIGURE 6. Run times for different (log) marginal likelihood estimators under various demographic priors. Log marginal likelihood estimation using PS/SS is markedly slower than both GSS implementations, across the demographic priors tested. The exponential-logistic model was omitted due to the PS/SS calculations failing for this model, leaving us without a basis for comparison in terms of the execution time. All estimators collected samples from 64 power posteriors that were run for 1 million iterations. The GSS approach using a matching coalescent model (MCM) yields the fastest run time for each demographic prior.

sampling close to the (diffuse) prior and resort to a change in likelihood scaling in BEAST/BEAGLE, GSS avoids this by collecting samples along a path from posterior to working distribution. Furthermore, providing a matching coalescent model (MCM) as a working distribution for the coalescent process reduces the computational burden compared with calculating a product of exponential distributions at every iteration, the complexity of which increases linearly with the number of taxa. As a consequence, the GSS MCM approach may represent the more convenient choice for comparing simple parametric demographic priors.

HIV-1 subtype C evolutionary patterns.—We analyze 81 HIV-1 subtype C sequences (Vrancken et al. 2014), consisting of seven genes (*gag*, *pol*, *env*, *vpr*, *vpu*, *vif*, and *nef*), using codon partitioned nucleotide substitution models. The data set consists of a diverse and representative (in terms of diversity) subset of all available HIV-1 subtype C full genomes with known sampling year from the Los Alamos HIV sequence database (http://www.hiv.lanl.gov/) and spans the period of 1986–2010. As in the original analysis, (Vrancken et al. 2014), we partition the full genome by gene, to allow for among-gene rate variation, and per gene by codon position as a trade-off between computational efficiency and biological realism (Shapiro et al. 2006; Baele and Lemey 2013).

These partitioning schemes offer some of the most popular approaches when analyzing coding data sets, and most often a choice is made between grouping the first and second codon positions together (the "112" notation, where one evolutionary model is used for the first and second codon positions and a second evolutionary model is used for the third codon position) or analyzing them separately (the "123" notation, where a separate evolutionary model is used for each of the three codon positions). We combine these two partitioning schemes with one of the most popular evolutionary models, that is the GTR model of nucleotide substitution (Tavaré 1986), and test the two different partitioning schemes using three (log) marginal likelihood estimators: PS, SS, and GSS. For both schemes, we assume a constant population size model with a Gamma(0.001, 0.001) prior on the population size parameter, a Normal(log(0.003), 2.0) prior and an exponential(1/3) prior on the mean and standard deviation respectively of the lognormal distribution of the uncorrelated relaxed clock, and a lognormal(0.0, 1.0) prior on the relative rate parameters for the codon position partitions.

To specifically assess the influence of different prior choices on the performances of the various (log) marginal likelihood estimators, we have tested both moderately and highly diffuse priors on the parameters of the GTR models. For the GTR model, this means that the r_{AC}, r_{AT}, r_{CG}, and r_{GT} parameters are equipped with a relatively diffuse Gamma(0.05, 0.10) prior, whereas the r_{AG} parameter receives a Gamma(0.05, 0.05) prior and the r_{CT} parameter is set to 1.0; their very diffuse counterparts come in the form of Gamma(0.005, 0.01) and Gamma(0.005, 0.005) priors. For the GTR112 model, this leads to 70 priors on the different rate parameters, whereas for the GTR123 model, 105 priors are being provided. To assess how "vague" or "diffuse" our prior choices are for our analyses of this data set, we have also estimated KL distances for all the parameters in our evolutionary models (see Supplementary Material available on Dryad (http://dx.doi.org/10.5061/dryad.8tm76).

We estimated log marginal likelihoods under PS, SS, and GSS using a 10 million posterior exploration for all estimators followed by sampling from 65 power posteriors, spread according to a Beta(0.3, 1.0) distribution. To evaluate convergence, each power posterior was explored using 500 thousand, 1 million, and 2 million iterations. For the GTR112 model and GTR123 models, equipped with moderately diffuse priors, exploring 1 million iterations per power posterior proved sufficient for all estimators. For the GTR112 model, PS and SS overestimated the log marginal likelihood by about 50 log units (PS: −137334.80; SS: −137339.40; GSS: −137390.90), whereas for the GTR123 model this overestimation amounted to about 80 log units (PS: −137168.03; SS: −137173.84; GSS: −137252.58) when compared with GSS. Differences of this magnitude can easily lead to differences in the outcome of the model selection process, particularly when analyzing highly partitioned data sets.

Each (log) marginal likelihood estimator is associated with some variance, but the observed differences between estimators are much larger than the variability for each estimator, indicating that the vague priors may contribute to this overestimation in PS/SS. As shown in the ILP results described earlier in this article, small test cases allow computation of the (log) marginal likelihood by drawing a large number of samples directly from the prior. PS and SS rely on constructing a Markov chain for each power posterior, and use MCMC to collect samples from these power posteriors. Although this approach remains relatively efficient in the presence of data, that is close to the posterior, it becomes more challenging to characterize densities close to the prior. In other words, integrating near very diffuse priors using MCMC with its dependent samples may be relatively inefficient given the vast volume of parameter space with nearly equal prior density.

When very diffuse priors are specified on the parameters of the GTR112 model and GTR123 models, convergence issues arise for all estimators albeit to a different extent. Exploring each of the 65 power posteriors for up to 2 million iterations, PS and SS yield continually increasing estimates of the log marginal likelihood, unable to stabilize to even report a range in which the log marginal likelihood can be captured. Moreover, the estimates appear to differ by several hundreds of log units compared with GSS. In other words, exploring such a large amount of very diffuse priors for a model that is likely to be severely overparameterized, does not lead to convergence for PS and SS in this case. The GSS estimate of the log marginal likelihood varies within a range of 100 log units, but the computational settings used here were insufficient to reach adequate convergence. This result may appear counterintuitive, given that working distributions are provided for all the parameters and GSS avoids the need to actually explore the priors. However, convergence issues arise even at the posterior, before any attempt to estimate marginal likelihoods, which may be attributed to the overparameterization and use of very diffuse priors. These priors were used to explore the estimator stability and are not recommendable, and we note that they differ from the moderately diffuse priors typically used with these models.

Discussion

Bayesian phylogenetics requires a sensible balance between parameter richness and biological realism. A good model captures the key features of the hypothesis under investigation without introducing unnecessary error, bias and over-fitting. Accurate model comparisons are therefore a crucial part of phylogenetic hypothesis testing, even though all evolutionary models necessarily oversimplify reality. Recent developments in marginal likelihood estimation, such as PS (Lartillot and Philippe 2006) and SS (Xie et al. 2011), demonstrate the potential for more accurate Bayesian model selection while accommodating uncertainty about the underlying time-measured genealogy. These approaches are finding applications in an increasing amount of phylogenetic studies because they have proven to outperform previously used marginal likelihood estimators. One point of criticism concerning PS and SS, however, is that they are computationally much more demanding than posterior-based marginal likelihood estimators, which only require samples from the posterior distribution to perform model selection and can hence be calculated from a standard MCMC run.

Because of faster convergence and lower estimation variance, GSS requires less computational effort to achieve the same accuracy as PS and SS. As with PS/SS, the accuracy of GSS improves with increasing computational investment, that is a larger number of power posteriors and a longer chain length per power posterior. These settings are dependent on the size of the data set being analyzed and on the complexity of the model, making it difficult to suggest general computational settings that guarantee convergence of the (log) marginal likelihood estimate. Based on our empirical results, we suggest using a(n initial) chain length per power posterior of 1 million iterations to ensure convergence for each power posterior. The number of power posteriors can initially be set to between 50 and 100. Varying both settings between different independent estimations is a good strategy to assess convergence.

Lartillot and Philippe (2006) note that the difference between the estimated logarithm of the marginal likelihoods of two phylogenetic models can be small compared with the actual log marginal likelihoods, which can lead to a poor estimate of the BF unless the precision on each marginal likelihood estimate is very high. To counter this effect, a single path connecting the two competing models in the space of unnormalized densities can be constructed and the BF can be calculated directly along this single path (Gelman and Meng 1998). By construction, this approach often results in lower estimation error for the BF in phylogenetics (Rodrigue et al. 2006; Baele et al. 2013a). The approach that we adopt here to ease the path integration is to shorten the path from posterior to prior while still calculating the marginal likelihood for each model separately. We follow the approach recently proposed by Fan et al. (2011) that involves introducing an arbitrary "working" distribution that, in practice, one specifies as a product of independent probability densities parameterized using MCMC samples from the posterior distribution. The method was however restricted to evaluations on a fixed phylogenetic tree topology, as integrating over plausible tree topologies complicates generalized SS because of the need to define a working distribution for topologies that provides a good approximation to the posterior. In this article, we provide two approaches to accommodate phylogenetic uncertainty into GSS. A first approach involves specifying a "working" distribution based on the coalescent tree prior, for example by parameterizing this model using its mean population size(s) and mean growth rate. A second approach borrows ideas from the Bayesian skyride model (Minin et al. 2008) and specifies a product of exponential densities as a genealogical working distribution. Both approaches are shown to outperform PS and SS in a large coalescent-based phylogenetic simulation study, with GSS POEL offering increased accuracy over GSS MCM in our analyses of an HIV-1 empirical data set. We have put online a tutorial on how to use GSS in BEAST: https://rega.kuleuven.be/cev/ecv/tutorials/. Note that we have only explored the GSS in particular genealogical scenarios, and that applications to more specific approaches (e.g., multispecies coalescent) still need to be investigated.

Arima and Tardella (2012) have proposed an alternative approach to estimate (log) marginal likelihoods in phylogenetics that offers promising results. Their generalized harmonic mean estimator (GHME) method requires an auxiliary probability density that approximates the posterior, which in principle yields a very efficient estimator when this density is set as close as possible to the posterior. Arima and Tardella (2012) propose using a set of working distributions, denoted $\pi_0(\theta|M)$ in their article, as the required auxiliary density, similar to what is used in the GSS approach. In order to accommodate phylogenetic uncertainty in their approach, the genealogical working distributions we propose may be interesting to further explore in that context. However, it still remains to be evaluated how well such a GHME would perform compared with other (log) marginal likelihood estimators.

Although state-of-the-art procedures such as PS and SS have been shown to achieve good accuracy in Bayesian phylogenetic model testing, the computational demand for complex models on relatively large data sets represents a significant challenge in marginal likelihood estimation. The GSS approaches we propose here yield higher accuracy for the same computational investment, or in other words, they can attain the same degree of accuracy with less computational demands. In addition, we have shown that using GSS protects against numerical difficulties and hence overestimating the marginal likelihood when specifying vague priors, as often employed phylogenetics. Future work will need to address how GSS stacks up in terms of accuracy against a direct Bayes Factor estimation approach, as proposed by Lartillot and Philippe (2006), which eliminates potential problems with sampling from the prior for common parameters in the models being compared.

Supplementary Material

Supplementary material can be found in the Dryad data repository at http://dx.doi.org/10.5061/dryad.8tm76.

Funding

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864 and the National Institutes of Health (R01 AI107034, R01 HG006139 and LM011827) and the National Science Foundation (IIS 1251151 and DMS 1264153). The National Evolutionary Synthesis Center (NESCent) catalyzed this collaboration through a working group (NSF EF-0423641). Guy Baele acknowledges support from a Research Grant of the Research Foundation—Flanders (FWO; Fonds Wetenschappelijk Onderzoek—Vlaanderen).

Acknowldgments

We would like to thank the editors, Frank Anderson and Mark Holder, as well as Paul Lewis and an anonymous reviewer for their constructive comments that helped improve this article.

References

Arima S., Tardella L. 2012. Improved harmonic mean estimator for phylogenetic model evidence. J. Comp. Biol. 19:418–438.

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

Baele G., Lemey P. 2014. Bayesian model selection in phylogenetics and genealogy-based population genetics. In Bernardo J. M., Bayarri M. J., Berger J. O., editors. Bayesian phylogenetics: methods, computational algorithms, and applications. Boca Raton, Florida: Chapman & Hall/CRC Mathematical & Computational Biology. pp. 55–90.

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

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

Baele G., Li W.L.S., Drummond A.J., Suchard M.A., Lemey P. 2013b. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. Mol. Biol. Evol. 30:239–243.

Bielejec F., Lemey P., Baele G., Rambaut A., Suchard M.A. 2014. πBUSS: a parallel beast/beagle utility for sequence simulation under complex evolutionary scenarios. BMC Bioinformatics 15:133.

Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Fan Y., Wu R., Chen M.H., Kuo L., Lewis P.O. 2011. Choosing among partition models in Bayesian phylogenetics. Mol. Biol. Evol. 28: 523–532.

Faria N.R., Rambaut A., Suchard M.A., Baele G., Bedford T., Ward M.J., Tatem A.J., Sousa J.D., Arinaminpathy N., Pépin J., Posada D., Peeters M., Pybus O.G., Lemey P. 2014. The early spread and epidemic ignition of hiv-1 in human populations. Science 346:56–61.

Friel N., Petitt A.N. 2008. Marginal likelihood estimation via power posteriors. J. R. Stat. Soc. B 70:589–607.

Gelman A., Meng X.-L. 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Stat. Sci. 13:163–185.

Gill M.S., Lemey P., Faria N.R., Rambaut A., Shapiro B., Suchard M.A. 2013. Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol. Biol. Evol. 30:713–724.

Gray R.R., Tatem A.J., Johnson J.A., Alekseyenko A.V., Pybus O.G., Suchard M.A., Salemi M. 2011. Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant staphylococcus aureus st239 genome-wide data within a Bayesian framework. Mol. Biol. Evol. 28:1593–1603.

Hasegawa M., Kishino H., Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. J. Mol. Evol. 22:160–174.

Holder M.T., Lewis P.O., Swofford D.L., Bryant D. 2014. Variable tree topology stepping-stone marginal likelihood estimation. In: Chen M.-H., Kuo L., Lewis P.O., editors. Bayesian phylogenetics: methods, algorithms, and applications. Vol. 1. New York: Chapman & Hall/CRC. pp. 95–111.

Huelsenbeck J.P., Ronquist F., Nielsen R., Bollback J.P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

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

Lartillot N., Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55:195–207.

Lepage T., Bryant D., Philippe H., Lartillot N. 2007. A general comparison of relaxed molecular clock models. Mol. Biol. Evol. 24:2669–2680.

Minin V.M., Bloomquist E.W., Suchard M.A. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 25:1459–1471.

Murphy K.P. 2007. Conjugate bayesian analysis of the gaussian distribution. Technical Report University of British Columbia.

Newton M.A., Raftery A.E. 1994. Approximating Bayesian inference with the weighted likelihood bootstrap. J. R. Stat. Soc. B 56:3–48.

Rodrigo A.G., Felsenstein J. 1999. Coalescent approaches to hiv population genetics. In: Crandall, K.A. editor, The Evolution of HIV. Baltimore, Maryland: Johns Hopkins University Press. pp. 233–272.

Rodrigue N., Philippe H., Lartillot N. 2006. Assessing site-interdependent phylogenetic models of sequence evolution. Mol. Biol. Evol. 23:1762–1775.

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

Shapiro B., Rambaut A., Drummond A.J. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 23:7–9.

Sheather S.J. 2004. Density estimation. Stat. Sci. 19:588–597.

Silverman B.W. 1986. The kernel method for univariate data. In: Density estimation for statistics and data analysis. London: Chapman & Hall/CRC, pp. 34–72

Sinsheimer J.S., Lake J.A., Little R.J. 1996. Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. Biometrics 52:193–210.

Steel M.A. 2005. Should phylogenetic models be trying to fit an elephant? Trends Genet. 21:307–309.

Suchard M.A., Weiss R.E., Sinsheimer J.S. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 18:1001–1013.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of dna sequences. In: Waterman M.S., editor. Some mathematical questions in biology: DNA sequence analysis. Providence, RI: American Mathematical Society, pp. 57–86.

Vrancken B., Rambaut A., Suchard M.A., Drummond A.J., Baele G., Derdelinckx I., Wijngaerden E.V., Vandamme A.-M., Laethem K.V., and Lemey P. 2014. The genealogical population dynamics of hiv-1 in a large transmission chain: bridging within and among host evolutionary rates. PLoS Comp. Biol. 10:e1003505.

Worobey, M., M. Gemmel, D. E. Teuwen, T. Haselkorn, K. Kunstman, M. Bunce, J. J. Muyembe, J. M. M. Kabongo, R. M. Kalengayi, E. V. Marck, M. Thomas, P. Gilbert, and S. M. Wolinsky. 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455:661–665.

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

Yang Z., Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–372.

Zhu T., Korber B.T., Nahmias A.J., Hooper E., Shaper P.M., Ho D.D. 1998. An african HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature 391:594–597.