Estimation of Speciation Times Under the Multispecies Coalescent

Jing Peng[1,*], David L. Swofford[2], and Laura Kubatko[1,3,4]

[1]Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

[2]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA

[3]Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, USA

[4]Mathematical Biosciences Institute, The Ohio State University, Columbus, OH 43210, USA

[*] *To whom correspondence should be addressed;*

E-mail: peng.650@osu.edu

Author to receive proofs:

Jing Peng
404 Cockins Hall
1958 Neil Ave.
Columbus, OH 43210
FAX: 614-292-2096
E-mail: peng.650@osu.edu

Abstract

Motivation: The coalescent model is now widely accepted as an effective model for incorporating variation in the evolutionary histories of individual genes into methods for phylogenetic inference from genome-scale data. However, because model-based analysis under the coalescent can be computationally expensive for large data sets, a variety of inferential frameworks and corresponding algorithms have been proposed for estimation of species-level phylogenies and the associated parameters, including the speciation times and effective population sizes.

Results: We consider the problem of estimating the timing of speciation events along a phylogeny in a coalescent framework. We propose a maximum *a posteriori* estimator based on composite likelihood (MAP_{CL}) for inferring these speciation times under a model of DNA sequence evolution for which exact site pattern probabilities can be computed. We demonstrate that the MAP_{CL} estimates are statistically consistent and asymptotically normally distributed, and we show how this result can be used to estimate their asymptotic variance. We also provide a more computationally efficient estimator of the asymptotic variance based on the nonparametric bootstrap. We evaluate the performance of our method using simulation and by application to an empirical dataset for gibbons.

Availability and implementation: The method has been implemented in the *PAUP** program, freely available at https://paup.phylosolutions.com for Macintosh, Windows, and Linux operating systems.

Contact: peng.650@osu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Though numerous methods have recently been developed for estimating species tree topologies, methods for estimating the associated speciation (divergence) times are less well-developed. Traditionally, divergence times have been obtained using maximum likelihood estimates of branch lengths from a concatenated alignment, but this approach has been shown to produce systematic errors because it fails to account for variation in gene genealogies and their associated gene divergence times. As a result, some node ages are overestimated while others are underestimated (Ogilvie *et al.*, 2017).

In contrast to concatenation, coalescent-based methods explicitly model variation in individual gene genealogies under the multispecies coalescent (MSC) model (Hudson, 1983; Rannala and Yang, 2003). Several widely used implementations provide estimates of either speciation times or internal branch lengths in addition to estimating the species tree topology. Of the methods that infer species trees from multilocus data using estimated gene trees ("summary statistic methods" or "summary methods"), *ASTRAL* (Sayyari and Mirarab, 2016) and *MP-EST* (Liu *et al.*, 2010) can also provide estimates of branch lengths. Branch-length estimates from both of these methods are technically statistically consistent (Liu *et al.*, 2010; Sayyari and Mirarab, 2016), but the property of consistency only holds when the input data consist of an unbiased sample of true gene trees. For empirical data, where gene trees must be estimated from sequence data, statistical consistency would only be achievable by using a statistically consistent method to infer gene trees while allowing the gene length to go to infinity (violating the MSC-model assumption of no recombination within genes). In fact, both *ASTRAL* and *MP-EST* have been shown to underestimate internal branch lengths (measured in coalescent units) when input gene tree estimation error increases (Sayyari and Mirarab, 2016). In addition, Yang (2002) showed that phylogenetic errors inflate the probability of incongruent gene trees and lead to biased estimation of internal branch lengths.

An alternative to summary methods is a fully Bayesian approach that jointly

estimates the species tree topology, speciation times, and effective population sizes using the complete sequence data (i.e., without first estimating gene trees for each locus). The most popular methods in this category are implemented in **BEAST/StarBEAST2* (Heled and Drummond, 2010; Ogilvie *et al.*, 2017) and *BPP* (Yang and Rannala, 2014; Rannala and Yang, 2017) for multilocus sequence data, and *SNAPP* (Bryant *et al.*, 2012) for biallelic SNP data. *StarBEAST2* and *BPP* differ in the prior distributions assumed for the species tree, the range of evolutionary models supported, and details of the MCMC strategies used to sample from the posterior distribution. Bayesian methods have the advantage of using all of the data, although estimates of branch lengths or node ages obtained by these methods may be sensitive to the choice of prior distributions, especially when the species are closely related and/or the sequences are very similar. Moreover, because they rely on MCMC, the computation is expensive for data sets with a large number of species and/or genes.

A third class of methods infers species trees directly from the sequence data without requiring separate estimation of gene trees for each locus. The most widely used example of this class, SVDQuartets (Chifman and Kubatko, 2014), is much faster than fully Bayesian approaches, but it can only estimate the topology of the species tree. Here we use some of the theory underlying SVDQuartets (Chifman and Kubatko, 2015) to derive an estimator for node ages under the MSC model along with the Jukes-Cantor (JC69) DNA substitution model (Jukes and Cantor, 1969), assuming a molecular clock. This estimator is not directly connected to SVDQuartets apart from being a quartet-based method that operates under the MSC assumptions. As such, it can be used to estimate speciation times on trees obtained using any method, although it is especially relevant for SVDQuartets, which does not intrinsically provide estimates of node ages or branch lengths.

Our proposed estimator differs from those described above in that it uses the posterior density based on the composite likelihood to obtain computationally efficient estimators in a model-based framework. Our method thus fills a gap between fast summary

methods that discard the sequence data after estimating gene trees, and fully Bayesian methods that integrate over gene trees but require computationally-intensive MCMC algorithms. By retaining the full data throughout, our estimator accommodates variability arising from both gene tree variation and the mutation process while remaining computationally efficient. We show that this estimator is statistically consistent and asymptotically normally distributed. Though the uncertainty in the estimator can be quantified by the theoretical asymptotic variance predicted by our normality result, we find that use of the nonparametric bootstrap provides a more accurate estimate of the variance of the estimates. The performance and computational cost associated with our method of speciation time estimation is compared with *BPP* using simulated datasets. We use a genome-scale dataset for gibbons (Carbone *et al.*, 2014; Veeramah *et al.*, 2015; Shi and Yang, 2018) to demonstrate the performance of our estimator for empirical data.

2 Methods

In a 4-taxon species tree, there are $4^4 = 256$ possible site patterns. Chifman and Kubatko (2015) show that each site pattern probability $p_{i_a i_b i_c i_d}$ on a 4-leaf species tree with species a, b ,c, and d for a specific observation $i_a i_b i_c i_d$, $i_j \in \{\text{A,C,G,T}\}$, at the tips of the tree can be written as a function of the effective population size θ and node ages $\boldsymbol{\tau}$ in the tree (in coalescent units) under the JC69 (Jukes and Cantor, 1969) model. Under this model as well as the molecular clock assumption, the rooted symmetric 4-leaf species tree $((a, b), (c, d))$ will have 9 distinct site pattern probabilities:

$$
\begin{aligned}
&p_{xxxx} && p_{xyxz} = p_{yxxz} = p_{xyzx} = p_{yxzx} \\
&p_{xxxy} = p_{xxyx} && p_{xxyz} \\
&p_{xyxx} = p_{yxxx} && p_{yzxx} \\
&p_{xyxy} = p_{yxxy} && p_{xyzw}, \\
&p_{xxyy}
\end{aligned}
$$

while the rooted asymmetric 4-leaf species tree $(a,(b,(c,d)))$ will have 11 distinct site pattern probabilities:

$$\begin{array}{lll} p_{xxxx} & p_{xyxy} = p_{yxxy} & p_{xxyz} \\ p_{xxxy} = p_{xxyx} & p_{xxyy} & p_{yzxx} \\ p_{xyxx} & p_{xyxz} = p_{xyzx} & p_{xyzw}, \\ p_{yxxx} & p_{yxxz} = p_{yxzx} & \end{array}$$

where x, y, z and w denote different nucleotides. For example, p_{xxxx} includes the site patterns p_{AAAA}, p_{CCCC}, p_{GGGG} and p_{TTTT}, which have identical probabilities under the model. As another example, p_{xxxy} includes the site patterns p_{AAAC}, p_{AAAG}, p_{AAAT}, p_{CCCA}, etc.

We use $\boldsymbol{p}^{\mathbf{S}} = \left(p_1^{\mathrm{S}}(\boldsymbol{\tau},\theta),\, p_2^{\mathrm{S}}(\boldsymbol{\tau},\theta),\, \ldots,\, p_9^{\mathrm{S}}(\boldsymbol{\tau},\theta)\right)$ to denote the 9 different site pattern probabilities arising from the symmetric 4-taxon species tree, augmenting the notation above to indicate the dependence of the site pattern probabilities on the quantities θ and $\boldsymbol{\tau}$. Likewise, $\boldsymbol{p}^{\mathbf{A}} = \left(p_1^{\mathrm{A}}(\boldsymbol{\tau},\theta),\, p_2^{\mathrm{A}}(\boldsymbol{\tau},\theta),\, \ldots,\, p_{11}^{\mathrm{A}}(\boldsymbol{\tau},\theta)\right)$ denotes the 11 distinct site pattern probabilities from the asymmetric 4-taxon species tree. In an alignment of length M, the site pattern frequencies for these classes can be modeled as a multinomial random variable under the assumption that the observed sites are independent, conditional on the species tree:

$$\boldsymbol{Z} \sim \begin{cases} \text{Multinomial}(M, \boldsymbol{p}^{\mathbf{S}}), & \text{for a symmetric tree;} \\ \text{Multinomial}(M, \boldsymbol{p}^{\mathbf{A}}), & \text{for an asymmetric tree,} \end{cases}$$

where $\boldsymbol{Z}$ is the vector of site pattern counts for the 9 or 11 distinct classes.

2.1 Maximum *a posteriori* estimation based on composite likelihood

2.1.1 The MAP_{CL} estimator

We can split a tree of arbitrary size into the subtrees induced by each quartet of four leaves, and write the likelihood of the observed site pattern frequencies for each quartet.

For example, in the 5-leaf species tree in Figure 1, we can consider all sets of 4 tips, to get $\binom{5}{4} = 5$ different quartets. For any quartet i, each site in an alignment of length M can be classified into one of n_i distinct site patterns, where n_i equals 11 if the quartet induces an asymmetric subtree of the full tree ($i \in \{1, 2\}$ in this case) or 9 if it induces a symmetric subtree ($i \in \{3, 4, 5\}$). For each site m, $m = 1, 2, \ldots, M$, and each quartet i, $i = 1, 2, \ldots, 5$, define $\boldsymbol{V_i^{(m)}}$ to be the random vector of length n_i that contains a 1 in the j^{th} entry if site pattern j is observed at that site and 0 in all other entries, and let $\boldsymbol{v_i^{(m)}} \in \{0, 1\}^{n_i \times 1}$ represent the corresponding observed data. Let $\boldsymbol{v_i} = \left(\boldsymbol{v_i^{(1)}}, \ldots, \boldsymbol{v_i^{(M)}}\right)$ denote the observed data across all M sites, and let $(u_i)_j$ be the j^{th} entry of the vector $\boldsymbol{u_i} = \sum_m \boldsymbol{v_i^{(m)}}$, which counts the number of times site pattern j is observed. The likelihood for quartet i can then be expressed as a function of the population size θ and node ages $\boldsymbol{\tau}$:

$$L_i(\boldsymbol{\tau}, \theta | \boldsymbol{u}_i) = \frac{M!}{\prod_{j=1}^{n_i} (u_i)_j!} \prod_{j=1}^{n_i} p_{ij}(\boldsymbol{\tau}, \theta)^{(u_i)_j}, \tag{1}$$

where p_{ij} is the j^{th} entry in either $\boldsymbol{p^S}$ or $\boldsymbol{p^A}$ for quartet i, depending on whether the subtree induced by this quartet is symmetric or asymmetric, respectively. Importantly, the subtrees induced by different quartets are not independent, and computing a true likelihood would require accounting for the correlation structure among quartets. Therefore, we instead use *composite likelihood*—the product of the individual likelihoods for all possible quartets despite their non-independence. Note that composite likelihood is also often referred to in the statistical and biological literature as *pseudolikelihood* or *approximate likelihood* (see Varin *et al.*, 2011, for a review of the history of composite likelihood methods).

A maximum composite likelihood estimator (MCLE) based on (1) would optimize the function

$$\ell(\boldsymbol{\tau}, \theta | \boldsymbol{x}) = \prod_{i=1}^{Q} L_i(\boldsymbol{\tau}, \theta | \boldsymbol{v}_i), \tag{2}$$

where Q is the number of possible quartets and the vector $\boldsymbol{x} = (\boldsymbol{x^{(1)}}, \cdots, \boldsymbol{x^{(M)}})$ is defined

similarly to the $\boldsymbol{v_i}$, but for the entire tree — i.e., its dimension depends on the number of possible distinct site patterns on a 5-leaf tree. Specifically, each vector $\boldsymbol{x^{(m)}}$ records which of the possible distinct site patterns on a tree of 5 tips is observed at site m, while the corresponding $\boldsymbol{v_i^{(m)}}$ stores the indicators of the n_i site patterns for the i^{th} quartet of this tree at site m.

Instead of using the MCLE, however, we prefer to estimate $\boldsymbol{\tau}$ and θ via Bayesian maximum *a posteriori* (MAP) estimation (e.g., Bassett and Deride, 2019). MAP estimation has two advantages. First, it allows incorporation of prior knowledge into the estimate as for the fully Bayesian methods discussed above. Perhaps more importantly, weighting the likelihood by the priors improves the computational efficiency and stability of the optimization algorithms by reducing the flatness of the optimality surface in regions of the parameter space that have very low likelihood.

With inclusion of the priors, the posterior density function becomes

$$g(\boldsymbol{\tau}, \theta|\boldsymbol{x}) = f_\theta(\theta)\ f_\mathrm{h}(\tau_\mathrm{R})\ \ell(\boldsymbol{\tau}, \theta|\boldsymbol{x}),$$

where f_θ and f_h are the prior density functions for θ and the height of the tree (= root age) τ_R, respectively. Non-root τ values are parameterized as proportions of the tree height, so that τ_R serves as a scaling factor for the entire tree. By maximizing the log posterior density

$$\log g(\boldsymbol{\tau}, \theta|\boldsymbol{x}) = \log f_\theta + \log f_\mathrm{h} + \sum_{i=1}^{Q} \log L_i(\boldsymbol{\tau}, \theta|\boldsymbol{v}_i), \tag{3}$$

we obtain our maximum *a posteriori* estimator MAP_CL:

$$(\tilde{\boldsymbol{\tau}}, \tilde{\theta}) = \underset{\boldsymbol{\tau}, \theta}{\operatorname{argmax}} \left\{\log g(\boldsymbol{\tau}, \theta|\boldsymbol{x})\right\}, \tag{4}$$

with the "CL" subscript signifying that a composite-likelihood term is used in (2) rather than a true likelihood.

2.1.2 Consistency and asymptotic variance calculation

Using a result from Arnold and Strauss (1991), we can prove that the MAP_{CL} estimator is statistically consistent and asymptotically normally distributed (a detailed proof can be found in the Supplemental Material, Section 1):

$$\sqrt{M}(\tilde{\tau}_k - \tau_k) \rightarrow N\left(0, \frac{K(\tau_k)}{J^2(\tau_k)}\right)$$

$$K(\tau_k) = \sum_{i,i'} \mathrm{E}_{\tau_k}\left[\left\{\frac{\partial}{\partial \tau_k} \log L_i(\boldsymbol{\tau}, \theta|\boldsymbol{v_i})\right\}\left\{\frac{\partial}{\partial \tau_k} \log L_{i'}(\boldsymbol{\tau}, \theta|\boldsymbol{v_{i'}})\right\}\right]$$

and

$$J(\tau_k) = -\sum_{i} \mathrm{E}_{\tau_k}\left[\frac{\partial^2}{\partial \tau_k^2} \log L_i(\boldsymbol{\tau}, \theta|\boldsymbol{v_i})\right],$$

where $\tilde{\tau}_k$ is the k^{th} component of the MAP_{CL} estimator $\tilde{\boldsymbol{\tau}}$.

This result provides a way of quantifying the uncertainty of our estimator, by calculating the first and second derivatives of the log likelihoods for the individual quartets. For example, to compute the asymptotic variance of the MAP_{CL} estimator $\tilde{\tau}_1$ for the 5-taxon species tree in Figure 1, note that quartets 3, 4, and 5 include node age τ_1. Therefore, if the relevant MAP_{CL} estimates are $\tilde{\boldsymbol{\tau}}$ and $\tilde{\theta}$, then we can approximate $K(\tau_1)$ and $J(\tau_1)$ by $K^*(\tau_1)$ and $J^*(\tau_1)$ as

$$K^*(\tau_1) = \sum_{i,i' \in \{3,4,5\}} \left[\left\{\frac{\partial}{\partial \tau_1} \log L_i(\boldsymbol{\tau}, \theta|\boldsymbol{v_i})\Big|_{\tilde{\boldsymbol{\tau}}, \tilde{\theta}}\right\} \left\{\frac{\partial}{\partial \tau_1} \log L_{i'}(\boldsymbol{\tau}, \theta|\boldsymbol{v_{i'}})\Big|_{\tilde{\boldsymbol{\tau}}, \tilde{\theta}}\right\}\right]$$

and

$$J^*(\tau_1) = \sum_{i \in \{3,4,5\}} \left\{\frac{\partial^2}{\partial \tau_1^2} \log L_i(\boldsymbol{\tau}, \theta|\boldsymbol{v_i})\Big|_{\tilde{\boldsymbol{\tau}}, \tilde{\theta}}\right\}.$$

The asymptotic variance is then calculated as

$$Var^*(\hat{\tau}_1) = \frac{K^*(\tau_1)}{J^*(\tau_1)^2} \ . \tag{5}$$

To calculate K^* and J^*, we plug in $\tilde{\boldsymbol{\tau}}$ and $\tilde{\theta}$ to estimate the 9 or 11 site pattern probabilities (rounding to 2 significant figures, which makes the variance estimates more stable).

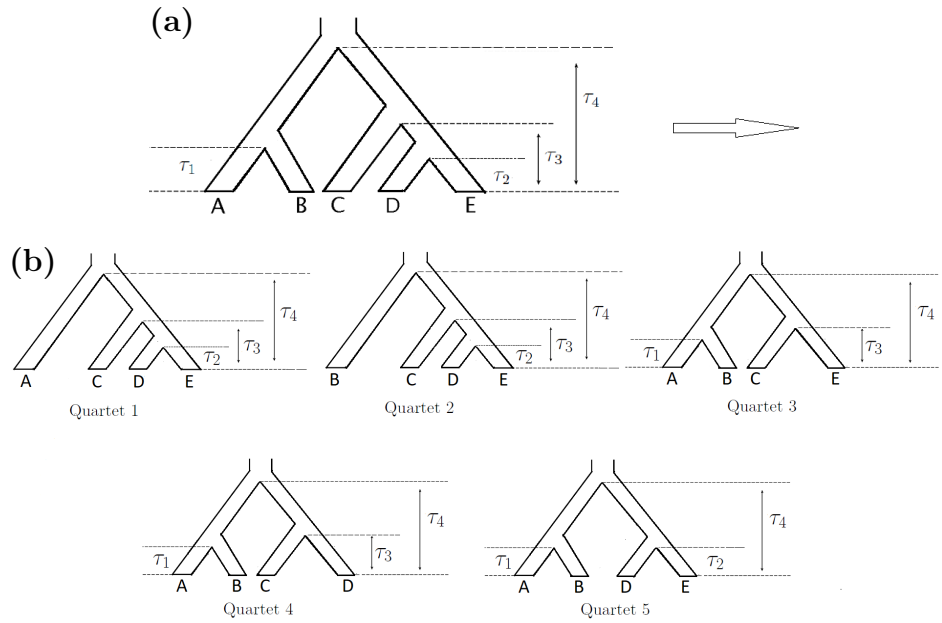


Fig. 1. The 5-leaf species tree (a) can be split into the 5 different 4-leaf subtrees (b), shown with speciation times marked.

We can also use a bootstrapping approach to estimate the variance of the MAP_{CL} estimator. In this approach, a bootstrap replicate is obtained by resampling the columns, i.e., site patterns in the original DNA sequences, using the following steps:

1. One column is randomly selected (with replacement) from the original sequence alignment to be a new column in the resampled data. By repeating this M times, we get a new bootstrap replicate of the same size as the original data;

2. Repeat step 1 to get B bootstrap samples;

3. For each of the B of the bootstrap samples, redo the analysis to compute the

estimates $(\tilde{\boldsymbol{\tau}}_1, \tilde{\theta}_1), \cdots, (\tilde{\boldsymbol{\tau}}_B, \tilde{\theta}_B)$, and calculate the sample variance of the estimates, $\mathrm{Var}(\tilde{\boldsymbol{\tau}}_B)$ and $\mathrm{Var}(\tilde{\theta}_B)$.

2.1.3 Implementation in *PAUP**

All of the methods described herein are implemented in the *PAUP** program written by DLS (https://paup.phylosolutions.com), where they are accessed using the *qAge* command (type `help qage;` at the command prompt for a description of the available options). A detailed explanation of the implementation, including parameterizations, mathematical details for likelihood and gradient evaluations, optimization strategies, and validation, is provided in the "Implementation of qAge in *PAUP**" document contained in the Supplemental Material.

2.2 Simulation study

2.2.1 Simulation 1: Statistical properties of the MAP_{CL} estimator.

We first use simulation to assess the statistical consistency and asymptotic normality of the MAP_{CL} estimator and to compare the two methods of measuring uncertainty (calculation of the theoretical asymptotic variance vs. bootstrapping). We note that while many methods for inferring species-level phylogenies are based on multilocus data, our method was originally designed for unlinked sites arising from the coalescent model, a data type that we call Coalescent Independent Sites (CIS). For multilocus data, all sites within a given locus are assumed to have evolved on the same genealogy and are not independent. Although we assume here that the site patterns in the sequences constitute independent draws from the distribution characterized by the MSC and nucleotide substitution models (Chifman and Kubatko, 2015), conditional on the species tree, a straightforward argument can be made that methods developed for CIS data can also be applied to multilocus data (Wascher and Kubatko, 2020), and we therefore consider both data types here.

To examine the properties of the MAP_{CL} estimator, we simulated two types of data:

(1) unlinked CIS data (each site evolves on its own own tree drawn randomly from the distribution of gene trees expected for the true simulation parameters under the MSC model), and (2) multilocus data (a sequence of length l is simulated for each locus on an underlying gene tree drawn randomly from the expected gene tree distribution). The simulations were performed as follows:

1. Generate gene tree samples under the MSC model based on a specified input species tree;

2. Generate DNA sequences of length l for each gene tree under the JC69 model ($l = 1$ for CIS data);

3. Choose prior distributions for the parameters;

4. Compute the site pattern frequencies for all possible quartets and maximize the log posterior density to obtain node age estimates using the MAP_{CL} estimator and estimate their theoretical asymptotic variances;

5. Resample the simulated sequences to get B bootstrap replicates, and compute the sample variance of the estimates via bootstrapping, as described in the previous section (for the multilocus datasets, a two-level bootstrap is conducted where we first take a bootstrap sample of genes followed by independent bootstrap resampling of sites within each gene);

6. Repeat steps 1–4 D times to obtain node age estimates and estimate variances using both theoretical asymptotic calculations and bootstrapping.

All steps in the simulations were carried out using the *qAge* command in *PAUP**. In step 1, we set up two different model species trees: a 5-leaf tree and a 6-leaf tree (Figure 2). Time is measured in coalescent units (number of generations scaled by $2N_e$, where N_e is the effective population size), and we set the population-scaled mutation rate to $\theta = 4N_e\mu = 0.002$ (constant throughout the tree), where μ is the mutation rate. For

speciation times, we assigned the vector $(\tau_1, \tau_2, \tau_3, \tau_4) = b \cdot (0.5, 0.5, 1.0, 1.5)$ for the 5-tip model tree and $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5) = b \cdot (0.5, 0.5, 1.0, 1.5, 2.0)$ for the 6-tip model tree. Different choices of b then involve stretching or shrinking the tree; any choice of b results in trees that satisfy the molecular clock. We considered three choices for b: $b = 1, 2$, and 4. In step 2, the gene length l was set to 1 for CIS data (i.e., we simulated 100,000 genealogies with one DNA site for each), while for multilocus data, we simulated 10,000 genes, each of length l=100. In step 3, we assigned diffuse inverse-gamma priors: $IG(3, 0.006)$ for θ and $IG(3, 2)$ for the age of root. In steps 5 and 6, we chose B=100 and D=100, respectively.

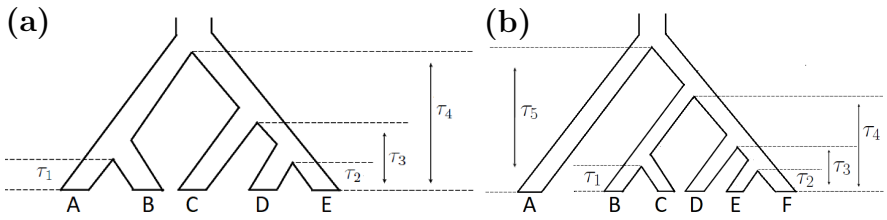


Fig. 2. Two different model species trees with speciation times as parameters for the simulation process: (a) 5-species tree. (b) 6-species tree.

The MAP_{CL} estimator is also applicable when multiple lineages are sampled for each tip species (see Supplemental Material: "Implementation"). To evaluate the performance of this option, the simulation framework within *PAUP** is instructed to generate gene tree samples using the same 5-leaf model species tree and parameter settings as above, but with two lineages for both species D and species E. We then use the *qAge* command in *PAUP** to repeat the analysis for the CIS data to obtain node ages and estimated variances.

2.2.2 Simulation 2: Comparison with *BPP*.

We carried out an additional simulation to compare the performance of the MAP_{CL} estimator in *qAge* with *BPP*, again using the simulation framework in *PAUP** (which provides an interface to invoke *BPP* from a Nexus file). We simulated multilocus data with 2,000 genes each of length 100, for trees with K tips (K=$7, 8, \cdots, 15, 20$). The trees used for simulation are included in the Supplemental Material Figure S16 and S17. The true

population-scaled mutation rate was set at $\theta = 0.002$. We ran *BPP* and *qAge* analyses with inverse-gamma priors $IG(\alpha, \beta)$ for θ and the age of the root node R (τ_{R}). Specifically, we used α=3 for a diffuse prior and adjusted β such that the prior mean $\beta/2$ was equal to 5θ, θ or $\theta/5$ for the θ parameter, and $5h$, h, or $h/5$ for the root age τ_{R}, where the tree height h is defined as the maximum number of branches connecting the tips and the root. For example, in Figure 2, the height of the tree is 3 in (a), and 4 in (b). The details of prior-distribution combinations can be found in Figure 6. We then investigated the impact of the 3×3 combinations of the priors on the performance of *BPP* and MAP_{CL}. To make the comparison in a computationally feasible way, for smaller trees (K=7, 8, 9, 10), we discarded the first 1,000 samples as burnin, and sampled every 50^{th} observation. For larger trees with more than 10 tips, we ran the *BPP* analysis 1,000 times longer than *qAge*, and discarded the first 10% samples as burnin. 500 observations were sampled equally frequently and used to compute estimates in all cases. The detailed MCMC configurations and running time can be found in the Supplemental Material (Table S1, Section 3). After performing analyses with *BPP* and *qAge* for 100 replicates, we quantified the deviation of estimated parameter values from the true values of the simulation model using the root-mean-square error (RMSE) and mean absolute error (MAE). We calculated the proportion of 95% confidence/credible intervals that included the true parameter value. Finally, for *BPP* analyses, we summarized the percentage of ESS values > 200.

2.3 Application to gibbon data

We also explored the performance of our MAP_{CL} estimator in inferring speciation times for empirical data by applying it to a genome-scale dataset previously analyzed by Shi and Yang (2018) for five species of gibbons: *Hylobates moloch* (Hm), *Hylobates pileatus* (Hp), *Nomascus leucogenys* (N), *Hoolock leuconedys* (B), and *Symphalangus syndactylus* (S) (Carbone *et al.*, 2014; Veeramah *et al.*, 2015). The dataset consists of 11,323 coding loci, each of length 200 bp. Except for the outgroup (O), multiple lineages are included for each

species: two for Hm and Hp, and four for N, B, and S. Here, we reanalyze these data with *qAge* (for MAP_{CL}) and *BPP*.

For both analyses, we fixed the species tree to be that shown in Figure 3. As recommended by the *BPP* authors (Flouri *et al.*, 2018), we use inverse-gamma prior distributions with the α parameter set to 3 for both θ and for the root age, τ_{R}. We then chose the value of β to match the mean of the distribution used by Shi and Yang (2018), although they assumed gamma rather than inverse-gamma prior distributions in an earlier version of *BPP*. To study sensitivity to the prior, we also conducted analyses with prior means that were five times larger and five times smaller than these values, and looked at all combinations of these prior settings for each parameter, leading to a total of nine prior combinations which we label Settings 1-9 in Table S2 of the Supplemental Material, Section 4 (see also Figure 8). Setting 5 corresponds most closely to the priors used in Shi and Yang (2018): $\theta \sim IG(3, 0.002)$ and $\tau_{\mathrm{R}} \sim IG(3, 0.02)$. For each choice of prior distribution, we repeated the analysis twice, with each replicate run for two weeks, and we sampled every 100^{th} observation. All prior settings reached at least 10,000 samples during this time (which corresponds to 10,000 $\times$ 100 = 1 million iterations of the algorithm), except for replicate 1 in setting 9, for which 9,205 samples were obtained. After discarding the first 2,000 samples as burnin, samples 2,000-10,000 from both replicates were combined to compute estimates (for setting 9, replicate 1, samples 2,000 - 9,205 were used).

For MAP_{CL}, we enumerated all possible quartets by selecting one lineage per tip for this tree, resulting in 752 quartet likelihoods used to calculate the composite likelihood. We assumed equal population sizes for all groups. Using the same priors as for *BPP*, we estimated the internal node ages ($\tau_{\mathrm{BS}}, \tau_{\mathrm{NBS}}, \tau_{\mathrm{HpHm}}$) and the population size (θ). Note that although we used the difference between speciation-time estimates in this case (i.e., branch lengths), statistical consistency still holds and asymptotic normality can be proved easily based on the asymptotically multivariate normal property for p-dimensional parameters given by Equation (2.7) in Arnold and Strauss (1991). The covariance matrix could be

obtained, but our simulations indicated that it may be unstable in comparison to bootstrap estimation. Therefore, 100 bootstrap samples were used to measure the uncertainty of our estimates.

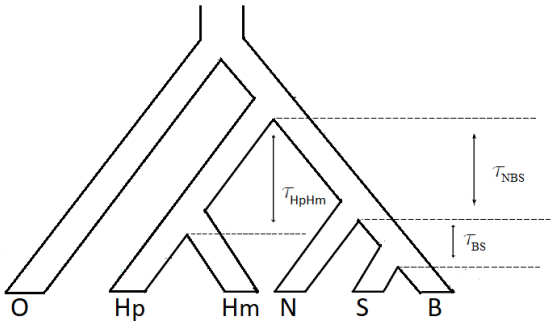


Fig. 3. The species tree for the five gibbon species and the outgroup (O=human) with node age parameters labeled.

3 Results

3.1 Simulation study

3.1.1 Simulation 1: Statistical properties of the MAP_{CL} estimator.

We plot histograms of the 100 MAP_{CL} estimates for node ages in the three 5-taxon and the three 6-taxon model trees (see the Supplemental Material, Section 2 for figures under all of the simulation settings). As a representative example, Figure 4 shows histograms of the 100 MAP_{CL} estimates of node age τ_1 for the three 5-leaf model trees under our simulation conditions. From these plots, we see that the estimates are approximately normal and distributed around the true value. Thus it appears that our estimates are unbiased. Moreover, when we include multiple lineages per tip or analyze multilocus data in the same way, the unbiasedness and asymptotic normality still hold, and if we increase the number of sites, we see these results even more clearly (see figures in the Supplemental Material, Section 2).

To assess the performance of our method in estimating the uncertainty of the MAP_{CL}

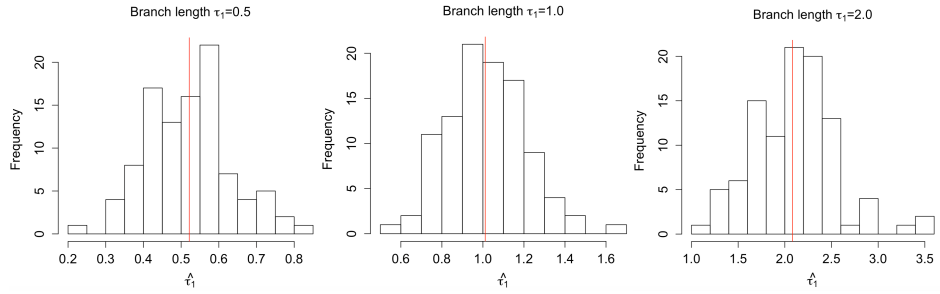
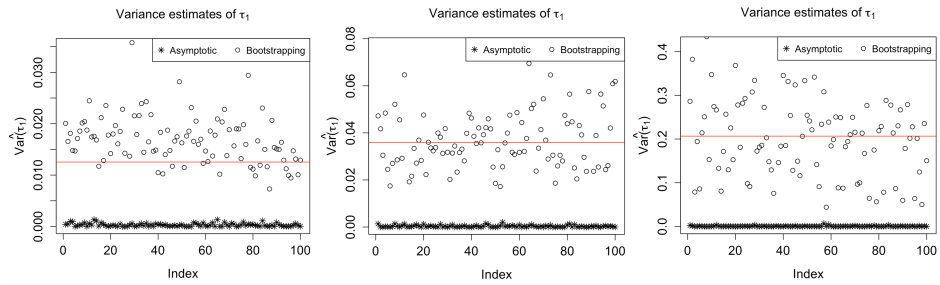


Fig. 4. Histograms of 100 MAP_{CL} estimates for node age τ_1 using 100,000 unlinked CIS from the 5-leaf model trees with a single lineage per tip. The red line in each histogram is the sample mean of the 100 MAP_{CL} estimates. The true values are given in the figure titles.

estimator, Figure 5 shows plots of the 100 variance estimates of the MAP_{CL} estimates of node age τ_1 for the three 5-leaf model trees under our simulation conditions. In the unlinked-CIS, single-lineage-per-tip setting, it is immediately clear that in all cases the bootstrap estimates are scattered around the sample variance from the 100 simulated datasets. This approximation is expected to improve as the number of sites increases. However, asymptotic variance estimates calculated by Equation (5) tend to underestimate the variance and to be unstable. We elaborate on this issue further in the Discussion, but note here that Varin *et al.* (2011) also remarked that, in practice, the bootstrap sometimes outperforms the asymptotic variance estimate in the composite likelihood setting. Thus, we recommend using the bootstrap estimator to measure the variance of the MAP_{CL} estimator, even though the asymptotic variance is theoretically reasonable. The results are similar when we use multilocus data or include multiple lineages per tip in the 5-taxon and 6-taxon model trees (see the Supplemental Material, Section 2).

3.1.2 Simulation 2: Comparison with *BPP*.

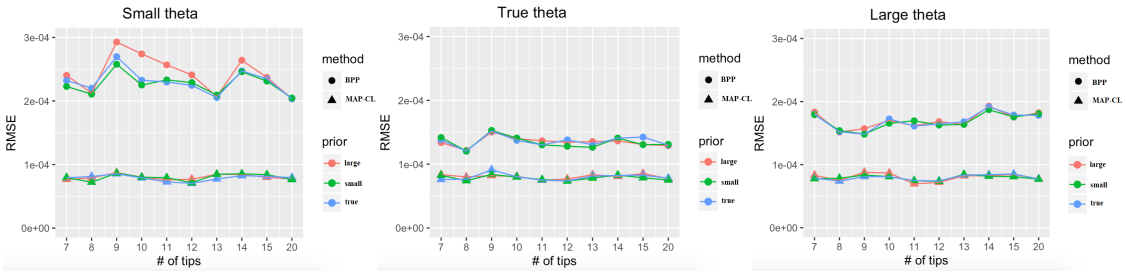
Next we compare our method with *BPP* and examine the estimation accuracy of both methods. Figure 6 summarizes the RMSE of the node age estimates on trees with different sizes. Overall, we find that MAP_{CL} estimates speciation times with smaller error than *BPP* and is quite robust to different prior combinations. The estimation error from *BPP* may be



$\tau_1 = 0.5$ $\tau_1 = 1.0$ $\tau_1 = 2.0$

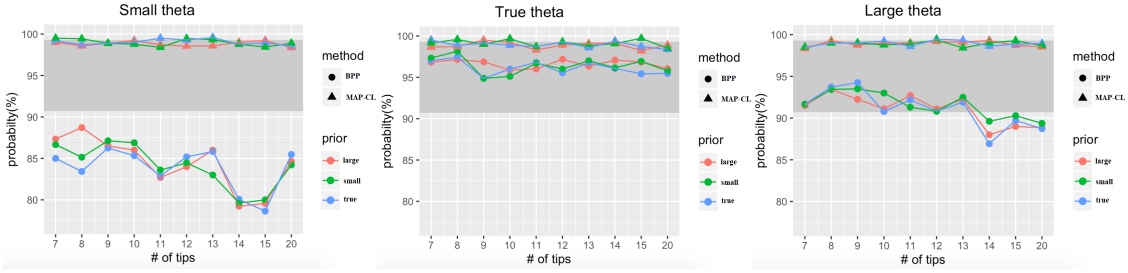
Fig. 5. Plots of the 100 variance estimates for node age τ_1 using 100,000 unlinked CIS from the 5-leaf model trees with a single lineage per tip. Points denoted by $*$ are computed by the asymptotic variance formula in (5), while points denoted by $\circ$ are obtained by bootstrapping. The x-axis is an index for the simulated samples. The red line in the plots is the sample variance of the 100 MAP_{CL} estimates.

partly due to convergence difficulties for some runs, which can be seen from the ESS values (see Figure S18). Moreover, the convergence problem can be ameliorated by choosing the prior for θ to have a large mean. Therefore, we conclude that after running *BPP* 1,000 times longer than *qAge*, our estimates are consistently comparable or more accurate than those from *BPP*. The results of MAE are similar to those for RMSE (Supplemental Material, Figure S17). Additionally, Figure 7 shows the proportion of 95% confidence/credible intervals that include the true parameter value in 100 simulation replicates. Again we note that the performance of *BPP* depends on the prior choice of θ, especially when we compare the coverage probabilities for large trees from (a) to (c), which generally do not show lack of convergence from the ESS values. On the other hand, the confidence intervals from MAP_{CL} include the true parameter values nearly 100% of the time, which highlights our finding that the bootstrap variance estimates overestimate the uncertainty when the number of genes is limited (Figure S7-S15; Section 2, Supplemental Material). We suggest that bias in this direction is acceptable, in that the resulting confidence intervals are conservative.



$\theta \sim IG(3, 0.0008)$ $\theta \sim IG(3, 0.004)$ $\theta \sim IG(3, 0.02)$

Fig. 6. Plots of the RMSE of the node age estimates for trees with varying numbers of tips. The x-axis shows the number of tips in the tree. Analysis based on two methods (circles – *BPP*, triangles – MAP_{CL}) is conducted with different priors. In each plot, the "large" prior for the root age, $\tau_{\mathrm{R}} \sim \mathrm{IG}(3,\ 5h)$, is shown in red; the "small" prior, $\tau_{\mathrm{R}} \sim \mathrm{IG}(3,\ h/5)$, is shown in green; and the prior centered at the true value, $\tau_{\mathrm{R}} \sim \mathrm{IG}(3,\ h)$, is shown in blue ($h$ is the tree height). Panels show the results of analyses using different priors on θ.



$\theta_{\mathrm{prior}} \sim IG(3, 0.0008)$ $\theta_{\mathrm{prior}} \sim IG(3, 0.004)$ $\theta_{\mathrm{prior}} \sim IG(3, 0.02)$

Fig. 7. Plots of the percentage of 95% confidence/credible intervals that include the true parameter value. The x-axis gives the tree size (number of tips). Points with different colors give values obtained using the 9 prior combinations for θ and τ_{R} (see also Figure 6). The shaded area gives the expected acceptance region of the coverage proportions in 100 simulation replicates.

3.2 Application to gibbon data

Results from *BPP* and MAP_{CL} are shown in Table 1 for the choice of prior distributions that corresponds most closely to those used by Shi and Yang (2018). For τ_{NBS} and τ_{BS}, the estimates from MAP_{CL} and *BPP* are similar, with wider confidence intervals for MAP_{CL} that cover the intervals given by *BPP*, as in the simulation studies. For τ_{HmHp}, however, the intervals given by MAP_{CL} and *BPP* do not overlap (though the values estimated are similar) and both are similar in width.

To examine the sensitivity of these estimators to the prior distribution, we evaluated both estimators under nine different prior settings, with the results shown in Figure 8. We

see that the estimates obtained using MAP_{CL} are robust to the choice of prior distribution, with little variation across the range of values selected. Conversely, *BPP* is sometimes strongly affected by the choice of priors, most notably for estimation of τ_{NBS} for settings 3 and 6. To examine this more carefully, we made trace plots of all parameters for all replicates and prior choices (see the Supplemental Material, Section 4). These trace plots show some cases in which the two replicates within a prior setting sampled different values for the entire run (see, e.g., the results for τ_{BS} in Figure S20 or for θ_{BS} in Figure S38, noting the difference in the y-axis values for setting 3; similar phenomena can be observed for the log likelihood – see Figure S65, setting 2 and Figure S66, setting 6). It is clear that for some settings, *BPP* experienced some difficulty converging, making clear that long runs may be required, even for the relatively straightforward problem of inferring node ages on a fixed 6-taxon species tree with a large data set. In contrast, *qAge* quickly produces stable MAP_{CL} estimates that are robust to the choice of prior distribution—it took only 17 seconds to produce an estimate and confidence interval for this data set on a current-generation laptop computer.

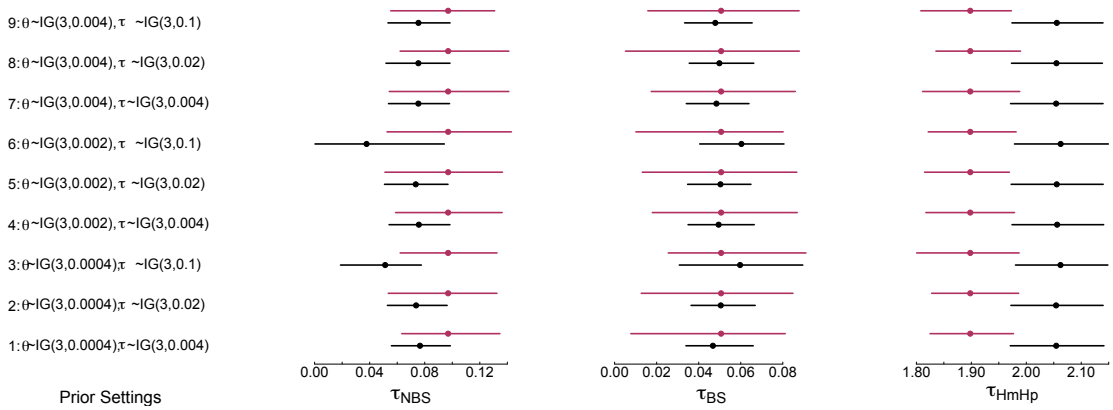


Fig. 8. 95% credible (*BPP*; black) and confidence (MAP_{CL}; maroon) intervals for the gibbon data for the 9 prior choices considered here (left panel). Setting 5 ($\theta \sim IG(3, 0.002)$ and $\tau_{\mathrm{R}} \sim IG(3, 0.02)$) is the closest match to the priors used by Shi and Yang (2018).

4 Discussion

4.1 Stability of the asymptotic variance

We note that in Figure 5 the calculated asymptotic variance from equation (5) is unstable and tends to underestimate the uncertainty, and this is not expected to improve by increasing the number of sites. As indicated in equation (5), derivatives must be computed to obtain the asymptotic variance; the first derivative of (1) can be expressed as:

$$\frac{\partial \log L_i(\boldsymbol{\tau}, \theta | \boldsymbol{u}_i)}{\partial \boldsymbol{\tau}} = \sum_{j=1}^{n_i} \left[\frac{(u_i)_j}{p_{ij}(\boldsymbol{\tau}, \theta)} \frac{\partial p_{ij}(\boldsymbol{\tau}, \theta)}{\partial \boldsymbol{\tau}} \right],$$

where p_{ij} is defined as in Equation 1.

Notice that the formulas above contain terms that include the reciprocal of site pattern probabilities, which are estimated by plugging in the MAP_{CL} estimates. Since most of these probabilities are very small (10^{-5}), even with estimation error on the order of 10^{-6}, the reciprocals vary by about 10^4. In fact, to get a good estimate, we need the relative error of the site pattern probabilities to be smaller than 0.001, but this is only around 0.01 in our case, even when the number of sites is very large. Therefore, though the asymptotic variance estimator is theoretically valid, it performs poorly in practice. We thus recommend using the bootstrap estimates to quantify the uncertainty of our estimator. This estimator is unbiased under the model assumptions, and it is quick to compute.

4.2 Computational efficiency of the MAP_{CL} estimator

Practically, to obtain MAP_{CL} estimates of the node ages on a species tree, we need to be able to do two things well: compute the composite likelihood, and search the parameter space to find the values that optimize the posterior probability density. The first of these can be done very efficiently for trees of arbitrary size. The number of individual likelihoods for all possible quartets (2) grows as the 4^{th} power of the tree size, but the amount of work

required per quartet is light, so that the total likelihood can be computed quickly even for a large tree.

The second task, however, becomes more difficult as the dimension of the parameter space increases. Fortunately, the gradient (first partial derivatives of the posterior density function with respect to each parameter) can be calculated quickly for any point in the parameter space, allowing the use of quasi-Newton and other gradient-based optimizers that typically need fewer function evaluations to converge to an optimum than derivative-free methods. In addition, the bootstrapping procedure used for measuring uncertainty could easily be parallelized, although we have not yet done so.

4.3 Assumptions and performance of the MAP_{CL} estimator

The assumptions that (1) nucleotide sites evolve according to the JC69 substitution model and (2) effective population sizes are constant throughout the tree, permit the use of formulas in Chifman and Kubatko (2015) for computing the 9 or 11 distinct site pattern probabilities used in Equation (1). Without these closed-form expressions, exact calculation of site pattern probabilities would involve an intractable multidimensional integration over gene trees and their associated branch lengths. Empirical data, however, may evolve under a nucleotide substitution model more complex than JC69, and preliminary simulations indicate that our method may not be robust when the nucleotide substitution model is misspecified and divergence between species is high (results not shown). However, for closely related species like gibbons, Shi and Yang (2018) argue that the JC69 model should be adequate for *BPP* and *ASTRAL*, and we note that *BPP* currently also assumes the JC69 model. Thus, the differences in the estimates obtained by *BPP* and *qAge* are not likely to be due to misspecification of the nucleotide substitution model. However, preliminary simulations indicate that our method may be somewhat sensitive to the assumption of constant effective population sizes across all populations, which may partially explain these differences.

Consequently, we are investigating plausible approaches for extending our method to allow inference under more general models, such as the GTR model and its submodels. An obvious, but computationally expensive, strategy would be to estimate site pattern probabilities by Monte Carlo simulation of a large number of independent sites under the assumed model for each point in parameter space visited by the optimizer. We are experimenting with an alternative method that makes a deterministic estimate of the desired vector of site pattern probabilities using the expected lengths of the branches on each possible gene tree, conditional on a species-tree topology and the current set of node-age and population-size parameter values. We expect future versions of *PAUP** to support this option (in a non-experimental mode), allowing the choice between using exact JC probabilities or an approximation of them using the branch-length expectation method under more complex models.

In the meantime, the MAP_{CL} estimator introduced in this paper is far more computationally efficient than fully Bayesian methods including *BPP* and *StarBEAST2*. With vague inverse-gamma priors, our initial exploration of prior sensitivity suggests that our estimates are robust to choice of prior means, given a reasonable sample size. *BPP*, on the other hand, may require a great deal of computation time to reach convergence when the prior and posterior distributions are centered around very different values. Unlike *ASTRAL* and *MP-EST*, our MAP_{CL} estimator does not require gene tree estimates (unbiased or otherwise), and it possesses the desirable properties of maximum likelihood and Bayesian estimators without requiring MCMC. The MAP_{CL} estimator can handle both CIS and multilocus data and can accommodate the sampling of multiple individuals per species. Finally, the guarantee of asymptotic normality ensures that the MAP_{CL} estimator will have good statistical properties as the amount of data increases. For these reasons, we anticipate that our MAP_{CL} estimator will be a useful addition to the collection of methods available for inferring speciation times from genome-scale data under the coalescent model.

Acknowledgments

We thank Jeff Thorne for helpful discussion regarding statistical issues, as well as the three anonymous reviewers who made important suggestions that strengthened the quality of the paper. We also acknowledge University of Florida Research Computing (http://rc.ufl.edu) and The Ohio State University College of Arts and Sciences (http://go.osu.edu/unitycompute) for providing computational resources.

Funding

This work was supported by the National Science Foundation [DEB 1455399 to L.K. and Andrea Wolfe; DMS 1610305 to L.K].

References

Arnold, B. C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhyā, Series B*, **53**, 233–243.

Bassett, R. and Deride, J. (2019). Maximum a posteriori estimators as a limit of Bayes estimators. *Math. Program.*, **174**(1-2), 129–144.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N., and RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.*, **29**(8), 1917–1932.

Carbone, L., Harris, R. A., Gnerre, S., Veeramah, K. R., Lorente-Galdos, B., Huddleston, J., Meyer, T. J., Herrero, J., Roos, C., and B. Aken et al (2014). Gibbon genome and the fast karyotype evolution of small apes. *Nature*, **513**, 195–201.

Chifman, J. and Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics*, **30**(23), 3317–3324.

Chifman, J. and Kubatko, L. (2015). Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *J. Theor. Biol.*, **374**, 35–47.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. (2018). Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.*, **35**(10), 2585–2593.

Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, **27**(3), 570–580.

Hudson, R. R. (1983). Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, **37**, 203–217.

Jukes, T. and Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–123. Academic Press, New York.

Liu, L., Yu, L., and Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.*, **10**(1), 302.

Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Mol. Biol. Evol.*, **34**, 2101–2114.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, **164**, 1645–1656.

Rannala, B. and Yang, Z. (2017). Efficient Bayesian species tree inference under the multispecies coalescent. *Syst. Biol.*, **66**, 823–842.

Sayyari, E. and Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. and Evol.*, **33**, 1654–1668.

Shi, C.-M. and Yang, Z. (2018). Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Mol. Biol. Evol.*, **35**, 159–179.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Stat. Sin.*, **21**, 5–42.

Veeramah, K. R., Woerner, A. E., Johnstone, L., Gut, I., Gut, M., Marques-Bonet, T., Carbone, L., Wall, J. D., and Hammer, M. F. (2015). Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate Bayesian computation approach. *Genetics*, **200**, 295–308.

Wascher, M. and Kubatko, L. (2020). Consistency of SVDQuartets and maximum likelihood for coalescent-based species tree estimation. *Syst. Biol., in press.*

Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, **162**, 1811–1823.

Yang, Z. and Rannala, B. (2014). Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.*, **31**, 3125–3135.

Table 1. Means and 95% credible (BPP) or confidence (MAP_{CL}) intervals in coalescent units for three internal node ages of interest for the gibbon dataset.[a]

Parameter	BPP		MAP_{CL}	
	mean	95% HPD interval	mean	95% CI
τ_{NBS}	0.07347	(0.05077, 0.09689)	0.09698	(0.05108, 0.13645)
τ_{BS}	0.05034	(0.03472, 0.06484)	0.05065	(0.01316, 0.08666)
τ_{HmHp}	2.05539	(1.97249, 2.13989)	1.89807	(1.81497, 1.96921)

[a] Using the prior distributions matching those used by Shi and Yang (2018) most closely: $\theta \sim IG(3, 0.002)$ and $\tau_{\mathrm{R}} \sim IG(3, 0.02)$ (our setting 5). For BPP, the conversion to coalescent units was carried out as in Shi and Yang (2018).