Systematic Biology Advance Access published November 18, 2010


DOI:10.1093/sysbio/syq073

# The Accuracy of Species Tree Estimation under Simulation: A Comparison of Methods

ADAM D. LEACHÉ* AND BRUCE RANNALA

*Genome Center and Department of Evolution and Ecology, University of California, Davis, CA 95616, USA;*
**Correspondence to be sent to: Department of Biology, University of Washington, Box 351800, Seattle, WA 98195, USA; E-mail: leache@uw.edu*

*Received 20 November 2009; reviews returned 14 March 2010; accepted 8 June 2010*
*Associate Editor: L. Lacey Knowles*

*Abstract.*—Numerous simulation studies have investigated the accuracy of phylogenetic inference of gene trees under maximum parsimony, maximum likelihood, and Bayesian techniques. The relative accuracy of species tree inference methods under simulation has received less study. The number of analytical techniques available for inferring species trees is increasing rapidly, and in this paper, we compare the performance of several species tree inference techniques at estimating recent species divergences using computer simulation. Simulating gene trees within species trees of different shapes and with varying tree lengths ($T$) and population sizes ($\theta$), and evolving sequences on those gene trees, allows us to determine how phylogenetic accuracy changes in relation to different levels of deep coalescence and phylogenetic signal. When the probability of discordance between the gene trees and the species tree is high (i.e., $T$ is small and/or $\theta$ is large), Bayesian species tree inference using the multispecies coalescent (BEST) outperforms other methods. The performance of all methods improves as the total length of the species tree is increased, which reflects the combined benefits of decreasing the probability of discordance between species trees and gene trees and gaining more accurate estimates for gene trees. Decreasing the probability of deep coalescences by reducing $\theta$ also leads to accuracy gains for most methods. Increasing the number of loci from 10 to 100 improves accuracy under difficult demographic scenarios (i.e., coalescent units $\leqslant 4N_e$), but 10 loci are adequate for estimating the correct species tree in cases where deep coalescence is limited or absent. In general, the correlation between the phylogenetic accuracy and the posterior probability values obtained from BEST is high, although posterior probabilities are overestimated when the prior distribution for $\theta$ is misspecified. [Coalescence; gene trees; incomplete lineage sorting; multilocus data; phylogeny reconstruction; simulation; tree shape.]

Simulation studies comparing the performance of phylogenetic inference techniques are essential for identifying situations where particular methods excel or perform poorly (Felsenstein 1978; Hillis et al. 1994; Hillis 1995). Recently, the estimation of species trees has become a central focus of systematic studies. A species trees is the multilocus estimate of the unobserved tree of genealogical relationships among species or populations as opposed to genealogies of single alleles. This increased focus on species tree inference is due in part to the ever-increasing ease of collecting multilocus data, a growing appreciation that gene tree variability renders individual genealogies unreliable predictors of the species tree, and the incorporation of the multispecies coalescent model in phylogenetic inference. New species tree inference techniques are emerging rapidly (reviewed by Edwards 2009), yet the relative accuracy of different approaches remains to be studied under a broad range of simulation conditions.

Modeling the processes that generate discordance between gene trees and species trees is the main objective of species tree inference (Pamilo and Nei 1988; Maddison 1997). One stochastic population-level process that can result in gene tree discordance is deep coalescence, which creates the opportunity for gene lineages to coalesce in ancestral populations in an order that does not match the species tree. Population demographics underlie the probability of observing deep coalescence events in gene genealogies, with the combination of large effective population sizes and short time intervals between speciation events producing the most discordance (reviewed by Degnan and Rosenberg 2009). Some species tree inference methods have incorporated the multispecies coalescent model to account for deep coalescence of gene lineages (e.g., Rannala and Yang 2003; reviewed by Liu et al. 2009), but the relationship between the expected levels of discordance between gene trees and species trees as a result of deep coalescence and the accurate estimation of the species tree is not well understood. Several studies examining the accuracy of species tree estimation using computer simulations have demonstrated that species tree inference methods may outperform methods based on data concatenation (Edwards et al. 2007; Kubatko and Degnan 2007; Liu et al. 2008; Kubatko et al. 2009) and that it is possible to infer species trees despite considerable incomplete lineage sorting (Maddison and Knowles 2006). Additional factors that can negatively impact the accurate estimation of the species tree include migration across population boundaries, sampling design, and gene tree estimation errors (Eckert and Carstens 2008; Huang and Knowles 2009; Leaché 2009; McCormack et al. 2009).

We conduct computer simulations to examine the relative accuracy of several species tree inference methods across a wide region of demographic parameter space that is representative of many empirical study systems. For example, the population sizes and tree lengths estimated for hominoids (Burgess and Yang 2008), *Drosophila* (Hey and Nielsen 2004), fence lizards (*Sceloporus*; Leaché 2009), and grass finches (*Poephila*; Jennings and Edwards 2005) are contained within the scope of demographic parameters that we use in our simulations. In general, previous simulation studies have contrasted species tree accuracy between relatively few simulation conditions (e.g., coalescent units $= 1N_e$ and $10N_e$; Maddison and Knowles 2006; McCormack

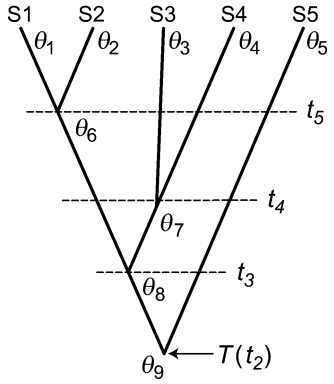


FIGURE 1. Species tree parameters illustrated on the maximally symmetric species tree, including the population size parameter theta ($\theta = 4Ne\mu$) and the total tree height ($T =$ generations $\times$ $\mu$). Estimating $\theta_1$–$\theta_5$is only possible when multiple samples are included for each species.

et al. 2009). Our simulations span a broad range of coalescent units ranging from $0.25N_e$ to $64N_e$ and also include many intermediate values. This simulation strategy increases the potential for overlap between the simulation conditions and empirical studies.

We compare the performance of phylogenetic methods that differ with respect to how gene tree incongruence is modeled, and this allows us to determine how effectively discordance between gene trees and species trees is accounted for by different methods using data simulated under the same demographic conditions. Because the probability of gene tree discordance also varies with tree shape (Tajima 1983; Degnan and Rosenberg 2006; Rosenberg and Tao 2008), we conduct simulations using asymmetric and symmetric species trees. By comparing phylogenetic accuracy between analyses utilizing subsets of the simulated data, we also address the question of how many loci may be necessary to obtain accurate estimates of the species trees under different demographic scenarios.

## METHODS

### *Species Tree Simulations*

We simulated rooted five-taxon species trees using the EVOLVER program in PAML version 4.1 (Yang 2007). The relative branch lengths for the species trees were generated under the birth–death process, with species sampling using a birth rate $\lambda = 3.0$, death rate $\mu = 0.2$, and sampling fraction $\rho = 0.2$ (Yang and Rannala 1997). These birth–death parameters resulted in a relatively uniform distribution of species trees node ages spanning a continuum with the following extremes: 1) node ages that are young relative to the root, producing trees having long internal branches and short tip branches and 2) node ages that are old and relatively close to the root, producing trees that are bush-like. This simulation procedure produced a large set of species trees with variable waiting times between speciation events (Figures S1 and S2, available from http://www.sysbio.oxfordjournals.org/). The molecular clock was assumed. We simulated 100 maximally symmetric trees $ST_{SYM} = (((1,2),(3,4)),5)$ (Fig. 1) and 100 asymmetric trees $ST_{ASYM} = ((((1,2),3),4),5)$ (Fig. 2) for each of the following tree heights, $T$ (the expected number of substitutions per site from the root to the tips): 0.001, 0.002, 0.004, 0.008, and 0.016. In total, we obtained 1000 species trees: 500 for each tree shape (asymmetric vs. symmetric) and 100 for each of the five values of $T$ (0.001, 0.002, 0.004, 0.008, and 0.016).

### *Gene Tree Simulations*

Within each species tree, we simulated gene trees and nucleotide sequence data using the MCcoal program in MCMCcoal version 1.2 (Yang 2009). The gene tree simulations accommodate deep coalescences by specifying the population size parameter (defined as $\theta = 4N_e\mu$, the product of the effective population size, $N_e$, and the per nucleotide site per generation mutation rate, $\mu$) on ancestral nodes (Rannala and Yang 2003). For any given tree height ($T$), increasing $\theta$ results in higher levels of deep coalescence and therefore the probability of simulating gene trees discordant with the species tree increases as well. While holding $\theta$ constant, decreasing $T$ results in the same pattern. For each species tree, we simulated 100 gene trees containing one sample per species using the following $\theta$ values: 0.001, 0.002, 0.004, 0.008, and 0.016 (Fig. 2). For each gene tree, we simulated 1000 base pairs of sequence data using the JC69 mutation model (Jukes and Cantor 1969). The model assumes no recombination within a locus, free recombination between loci, no migration between species, and neutral evolution. Thus, we simulated 100,000 bp of sequence data for each species tree, with 1 kb contributed from each of 100 loci. The simulated data are available online at http://www.sysbio.oxfordjournals.org/.

The 25 combinations of parameter values used for $T$ and $\theta$ to simulate data can be expressed on a timescale of $N_e$ generations using coalescent units (Kingman 1982, 2000). In general, the expected time for a large sample of alleles at a neutral locus to find their most recent common ancestor is $4N_e$ generations (Tajima 1983). The coalescent units for the species trees used in our study range from 0.25 (a high probability of deep coalescence) to 64 (a low probability of deep coalescence) and are calculated by dividing $T$ by $\theta/4$ (Fig. 2). The fact that species trees of the same coalescent units can arise from different combinations of population sizes or tree lengths is illustrated in Figure 2, where species trees with $4N_e$ coalescent units are represented by five separate parameterizations of $T$ and $\theta$. Although the coalescent processes are equivalent, the expected amount of information available from DNA substitutions is different under the various parameterizations. For example, the expected number of substitutions from the root of the tree to any tip is proportional to the absolute values of $T$ and $\theta$ (see Table 1).

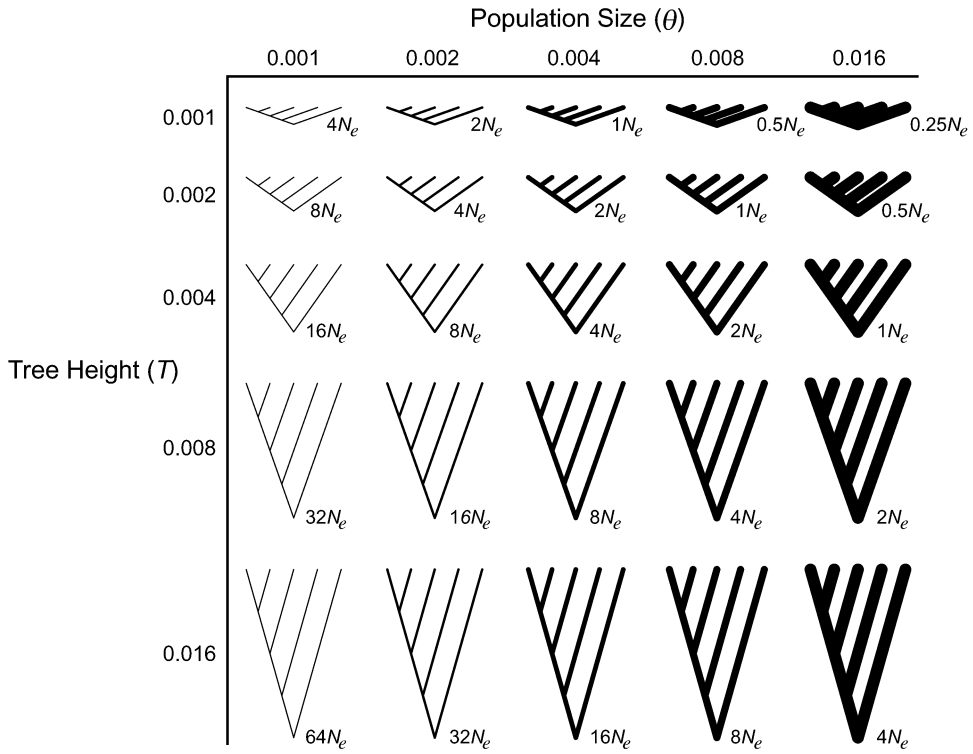


FIGURE 2. Combinations of population sizes ($\theta = 4N_e\mu$) and tree heights ($T$ = generations $\times$ $\mu$) used to simulate gene trees within species trees. Only asymmetric trees are shown, although simulations are also performed on maximally symmetric species trees $ST_{SYM} = ((1,2),(3,4)),5)$. Coalescent time units (time units normalized by population size) are shown next to each species tree and are obtained by dividing $T$ by $\theta/4$. In demographic terms, if $T$ and $\theta = 0.001$ and a mutation rate of $1 \times 10^{-8}$ is assumed, the populations size ($N_e$) is 25,000 (diploid individuals = 12,500) and the number of generations is 100,000. Decreasing the mutation rate by an order of magnitude to $1 \times 10^{-9}$ results in $N_e = 250{,}000$ and generations = 1,000,000.

### *Species Tree Inference*

We compared the accuracy of species tree estimation across three types of inference methods that differ with respect to how gene tree variability is accommodated. These methods (described below) included data concatenation, Bayesian concordance analysis, and models that incorporate the multispecies coalescent (maximum likelihood [ML] using gene trees and Bayesian inference). Phylogenetic accuracy, as used here, is measured as the percentage of times that a method obtained the true species tree, and when an estimated species tree contained a polytomy, we calculated the probability of the correct resolution. We measured accuracy on unrooted trees for comparisons among methods, which reduces the set of possible species trees from 105 to 15. We inferred species trees using 100,000 bp of data distributed across 100 independent loci (1 kb per locus). To examine the influence of the number of loci on accuracy, we also conducted species tree searches on reduced data sets containing only 10% of the simulated data (i.e., 10,000 bp or 10 loci).

*Concatenation.*—Data concatenation remain one of the most commonly used methods to analyze multilocus data despite the fact that this approach does not attempt to model gene tree variability resulting from deep coalescence. Instead, data from independent loci are combined into a "supermatrix" and analyzed as if they

TABLE 1. The average number of variable sites per 1 kb of sequence data simulated under different values of θ and $T$

| | | Population size (θ) | | | | |
|---|---|---|---|---|---|---|
| | | 0.001 | 0.002 | 0.004 | 0.008 | 0.016 |
| Tree height ($T$) | 0.001 | 5.66 ± 0.51 | 7.77 ± 0.62 | 12.00 ± 0.80 | 19.96 ± 1.14 | 35.45 ± 1.89 |
| | 0.002 | 9.12 ± 1.03 | 11.28 ± 1.04 | 15.43 ± 1.14 | 23.53 ± 1.45 | 39.68 ± 2.17 |
| | 0.004 | 16.01 ± 1.97 | 18.20 ± 1.97 | 22.35 ± 1.9 | 30.46 ± 1.91 | 46.43 ± 2.68 |
| | 0.008 | 29.53 ± 3.77 | 31.62 ± 3.78 | 36.05 ± 3.77 | 44.27 ± 3.9 | 60.08 ± 4.15 |
| | 0.016 | 55.78 ± 7.36 | 57.95 ± 7.30 | 62.25 ± 7.27 | 70.75 ± 7.09 | 86.92 ± 7.48 |

Notes: Standard deviations are shown below averages. Sequences were simulated under the JC model (Jukes and Cantor 1969) on gene trees generated from within the asymmetric species trees $ST_{ASYM} = ((((1,2),3),4),5)$.

represented one large contiguous locus. We concatenated the 100 loci (1 kb each) simulated for each species tree to produce supermatrices containing 100,000 bp. We inferred phylogeny for each supermatrix using maximum parsimony (MP) and Bayesian inference.

We conducted MP searches with the branch-and-bound search algorithm using PAUP version 4.0b1 (Swofford 2001). For the Bayesian analyses, we use MrBayes version 3.1.2 (Ronquist and Huelsenbeck 2003). We assumed the JC69 model of nucleotide substitution, which corresponds to the model used to simulate the data. Because some species tree inference methods included in our comparison inferred rooted trees (e.g., BEST and STEM), we enforced a molecular clock (assuming a uniform Dirichlet prior probability distribution on relative branch lengths) to increase the comparability between methods. An exponential ($\lambda = 1000$) prior distribution specified the tree height, which corresponds to a mean of 0.001 (i.e., $1/\lambda$) substitutions on a single branch extending from the root of the tree to the tips. As a result of concatenating large numbers of characters for just five species, convergence was reached quickly in the Bayesian analyses as determined by cumulative posterior probability burn-in plots constructed using the program Are We There Yet? (Nylander et al. 2008) produced for a subset of the results. We implemented Bayesian tree searches for 500,000 generations (sampling every 500 and discarding the first 500 samples as burn-in) using four concurrent chains (i.e., nchains = 4) with default heating values.

*Bayesian concordance analysis.*—The overlap among the posterior probability distributions for gene trees inferred from independent loci contains information related to the degree of concordance across those genes, and the predominant genealogical signal contained in these data is termed the primary concordance tree (Baum 2007). The primary concordance tree can be thought of as an estimate of the species tree that is built from those clades that have the highest posterior probability across the majority of the genome. We estimated primary concordance trees using the two-stage Bayesian concordance analysis method outlined by Ané et al. (2007). First, posterior probability distributions of gene trees were obtained for each locus using MrBayes version 3.1.2 under conditions similar to those used for the concatenated data analyses (see above). Second, an MCMC analysis implemented in BUCKy version 1.2b (Ané et al. 2007) was used to estimate the primary concordance tree from the posterior probability distributions obtained for the separate loci. Each BUCKy analysis utilized MCMC sampling with 100,000 generations (two independent analyses with default run convergence diagnostics), four chains per run, and a 10% burn-in factor. The Dirichlet process prior that controls gene tree clustering was set to $\alpha = 0.05$ to place high prior density on one underlying tree since the simulated data pertaining to any particular Bayesian concordance analysis were simulated from a single species tree as opposed to representing a collection of loci simulated from different underlying species histories.

*Coalescent models.*—The multispecies coalescent model accommodates deep coalescence into the phylogenetic inference of species trees (Rannala and Yang 2003; Liu et al. 2009). ML and Bayesian implementations of this model for species tree inference are available, and we use both here.

To obtain ML estimates of the species tree, we used the program STEM version 1.1a (Kubatko et al. 2009). Calculating the ML, species tree requires a set of resolved gene trees with branch lengths estimated from multiple loci and a point estimate for $\theta$. We inferred the ML point estimates for the gene trees using the branch-and-bound search algorithm in PAUP version 4.0b1. We assumed the JC model of nucleotide substitution and enforced a molecular clock. For each STEM analysis, we set $\theta$ to match the same value used to simulate the data. We used the simulated annealing algorithm (using default conditions) to calculate the likelihood score for each possible species tree topology. This was necessary because there were many instances where multiple species trees were tied for the ML score. In cases where the ML score for the correct species tree was tied with incorrect species trees, we calculated accuracy as the probability of randomly selecting the true tree out of the set of trees with equal ML scores. For example, if STEM recovered three species trees tied for the ML score, the probability of selecting the true species tree (if it is contained in the set of three trees) is $1/3$. This is analogous to the way we treated the other methods of analysis when an estimated species tree produced a trichotomy—the probability of selecting the true species tree (if it is one of the three possible resolved trees) is $1/3$.

For Bayesian inference of species trees, we use the hierarchical Bayesian model implemented in BEST version 2.2 (Liu 2008; Liu et al. 2008). The Bayesian approach estimates the joint posterior distribution of gene trees from unlinked loci and assumes that loci are correlated by a shared species history (Liu and Pearl 2007). The Bayesian method estimates the species tree directly from the sequence data and incorporates uncertainty associated with nucleotide substitution model parameter estimates, gene tree estimation, and the coalescent process (Liu 2008). Convergence using the joint prior is slow compared to the concatenation approach and requires longer MCMC analyses (Liu et al. 2008). We ran MCMC analyses for 10 million generations (sampling every 10,000 steps) with a 50% burn-in fraction. Convergence was assessed using burn-in plots of likelihood values and posterior probability values for clades from a subset of analyses. The prior distribution for the mutation rates across loci was set at uniform (0.1, 2.5), and the prior distribution for $\theta$ was modeled using an inverse gamma distribution ($\alpha = 3, \beta = 0.03$). The mean of the inverse gamma distribution is $\beta/(\alpha - 1)$, which corresponds to a prior mean for the population size of $\theta = 0.015$. We score accuracy using the 50% majority rule

consensus tree, and when the estimated species tree contained a polytomy, we calculated the probability of the correct resolution.

### *Credibility Intervals and Posterior Probability Accuracy*

Species tree estimation using the Bayesian implementation of the multispecies coalescent model (BEST) produces a posterior probability distribution for the species tree and posterior probability values for species relationships. Using our simulation results, we examined the accuracy of these posterior distributions on the rooted species trees. We expect that the size of the 95% credibility intervals will fluctuate with the levels of deep coalescence and that high levels of deep coalescence will result in large credibility intervals of species trees. We also examine the percentage of times that the true tree is contained in the 95% credible set, which evaluates the true coverage probability versus the nominal coverage probability. Finally, because the posterior probability values estimated for clades are commonly used measures of support for assessing confidence in phylogenetic analyses, we used our simulation results to determine the relationship between estimated posterior probability values and true phylogenetic accuracy.

## RESULTS

### *Expected Concordance of the Simulated Gene Trees*

In the absence of any deep coalescence events producing gene trees that are discordant with the species tree, we would expect to see 100% concordance between the simulated gene trees and species trees. However, the population size ($\theta$) and the tree height ($T$) parameters applied to the starting species trees result in variable levels of discordance across the simulated gene trees and species trees. We use the term "expected concordance" to refer to the concordance between the simulated gene trees and species trees. For the asymmetric species tree topology, the expected concordance is highest (86%) when the coalescent units $= 64N_e$ (Fig. 3). The expected concordance decreases as the coalescent units for the species trees decrease, and at $4N_e$ less than 25% of the simulated gene trees match the species tree (Fig. 3). The symmetric species trees produces less discordance overall, and the expected concordance ranges from a high of 95% ($64N_e$) to roughly 35% at $4N_e$ (Fig. 3).

### *Species Tree Accuracy*

Nucleotide sequence data simulated on the gene trees under the JC69 model contained 5.6–86.9 variable sites per 1 kb of sequence data (Table 1). Increasing either the species tree height ($T$) or the population size ($\theta$) resulted in an increase in the number of variable sites (Table 1). More substitutions are expected to occur on longer trees, and these data show that increasing $\theta$ results in deeper coalescence times for the gene trees (on average) and therefore also increases the number of variable sites.

The accuracy of species tree inference in response to 25 different parameterizations of $\theta$ and $T$ using data concatenation (parsimony and Bayesian inference), Bayesian concordance analysis (BUCKy), and two implementations of the multispecies coalescent model (ML using gene trees: STEM; Bayesian estimation: BEST) is shown in Figure 3. In general, each method of analysis results in a similar wave-like pattern, whereby species tree accuracy increases with both increasing values of $T$ and decreasing values of $\theta$ (Fig. 3). The STEM results are an exception to this pattern since accuracy is improved primarily by increases in $T$ (Fig. 3). Most of the inference procedures result in improved performance compared to the expected concordance (e.g., the accuracy that describes the concordance between the simulated gene trees and the species trees), and the overall accuracy for the symmetric trees is generally higher than the asymmetric trees (Fig. 3). This latter result may reflect an influence of the priors on the different tree shapes. The asymmetric species tree has one possible labeled history, whereas the maximally symmetric species tree has two and therefore has twice as much weight in the prior distribution.

Increasing the number of loci by 10-fold (10–100 loci) has variable results for the different analytical techniques. For the asymmetric species trees, the greatest increase in accuracy is seen with the multispecies coalescent model implemented in BEST, which outperforms the other methods under the most difficult demographic scenarios tested (small $T$ and large $\theta$; Fig. 3a). For the symmetric species tree, all methods have high accuracy ($\geqslant$ 95%) with 10 loci when $T$ is large and $\theta$ is small, which leaves little room for improvement when increasing the number of loci by 10-fold (Fig. 3b). Only STEM does not perform better with the addition of more loci, and accuracy decreases as much as 25% are seen on the symmetric species tree simulations (Fig. 3b).

### *Comparison of Methods*

Direct comparisons of the performance of the species tree inference methods across the 25 different parameterizations of $\theta$ and $T$ are shown in Figure 4. All methods perform well at low values of $\theta$ and high values of $T$ (Fig. 4). However, under the more challenging conditions of a high $\theta$ and low $T$, BEST outperforms the other methods (Fig. 4). Data concatenation using MP or Bayesian inference performs equally well for the asymmetric species trees, but Bayesian inference shows a slight increase in accuracy under most simulation conditions on the symmetric species trees (Fig. 4). The BEST method outperforms BUCKy and data concatenation under a wide range of simulation conditions on the asymmetric species trees (Fig. 4a). For the symmetric species trees, BUCKy and BEST outperform data concatenation under most simulation conditions, while BUCKy is often slightly more accurate than BEST (Fig. 4b).

The performance of STEM is generally lower compared to the other methods, and accuracy is often lower

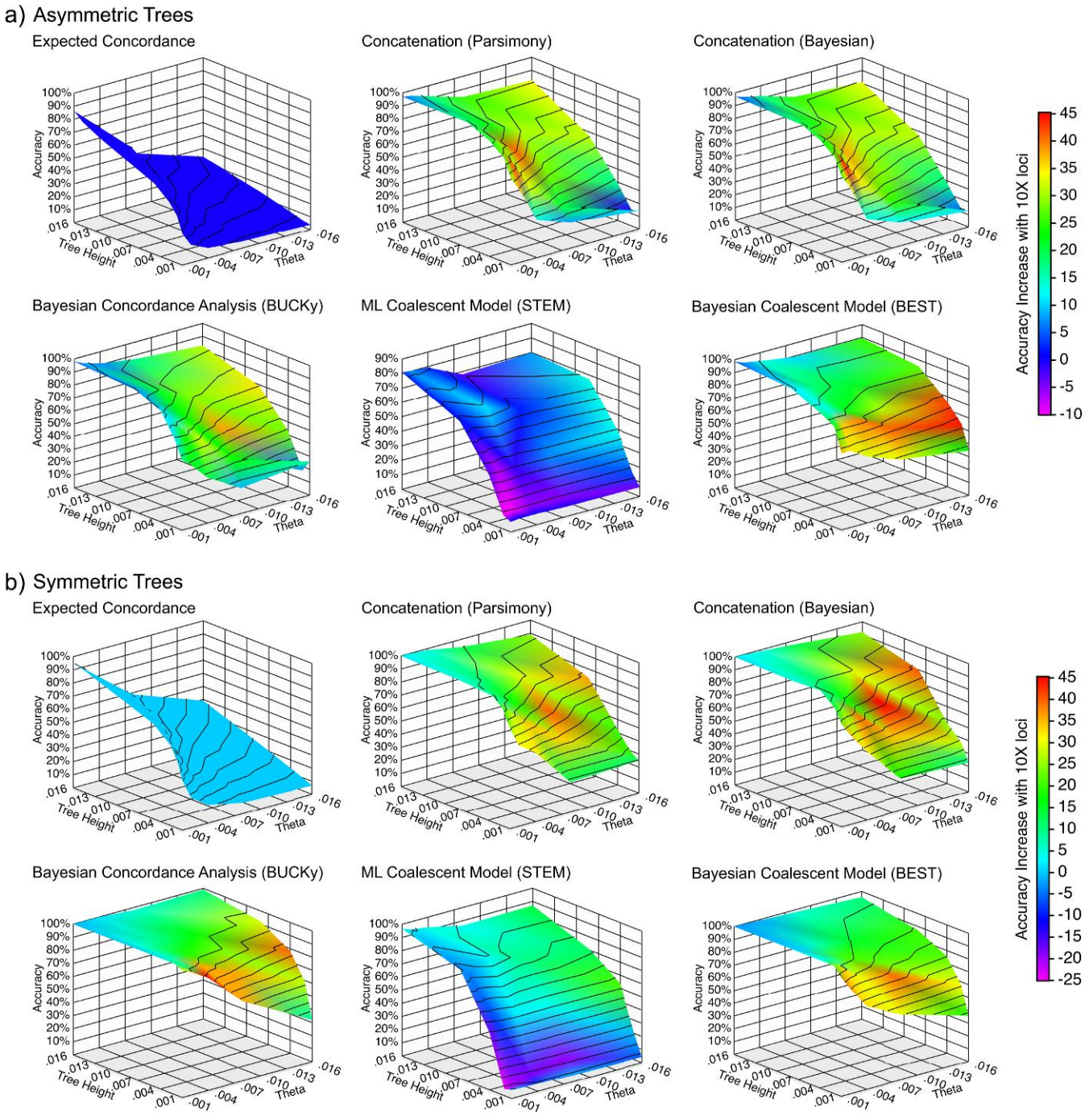


FIGURE 3. Accuracy of species tree inference methods for the (i) asymmetric and the (ii) symmetric species trees using 100 loci. Species tree accuracy is plotted on the *z*-axis in relation to theta (θ; *x*-axis) and tree height (*T*; *y*-axis). The improvement in species tree accuracy as a result of increasing the number of loci by 10-fold (10–100 loci) is illustrated in color on each contour plot and is standardized across the different analytical methods.

than that of the expected concordance. This result is expected because gene tree inference errors, in addition to the coalescent process, are contributing to species tree uncertainty (Huang and Knowles 2009). Increasing the amount of information available in the sequence data to accurately infer gene trees provides STEM with dramatic gains in accuracy, and this is accomplished primarily by increasing *T* (Fig. 4). In contrast to the other methods of analysis, changes in θ do not have any clear impact on accuracy when using STEM (Fig. 4). One source for the low accuracy of STEM results from a lack of information regarding the ML estimate of the species tree (i.e., many species trees are tied for the ML score) rather than the method being biased for an

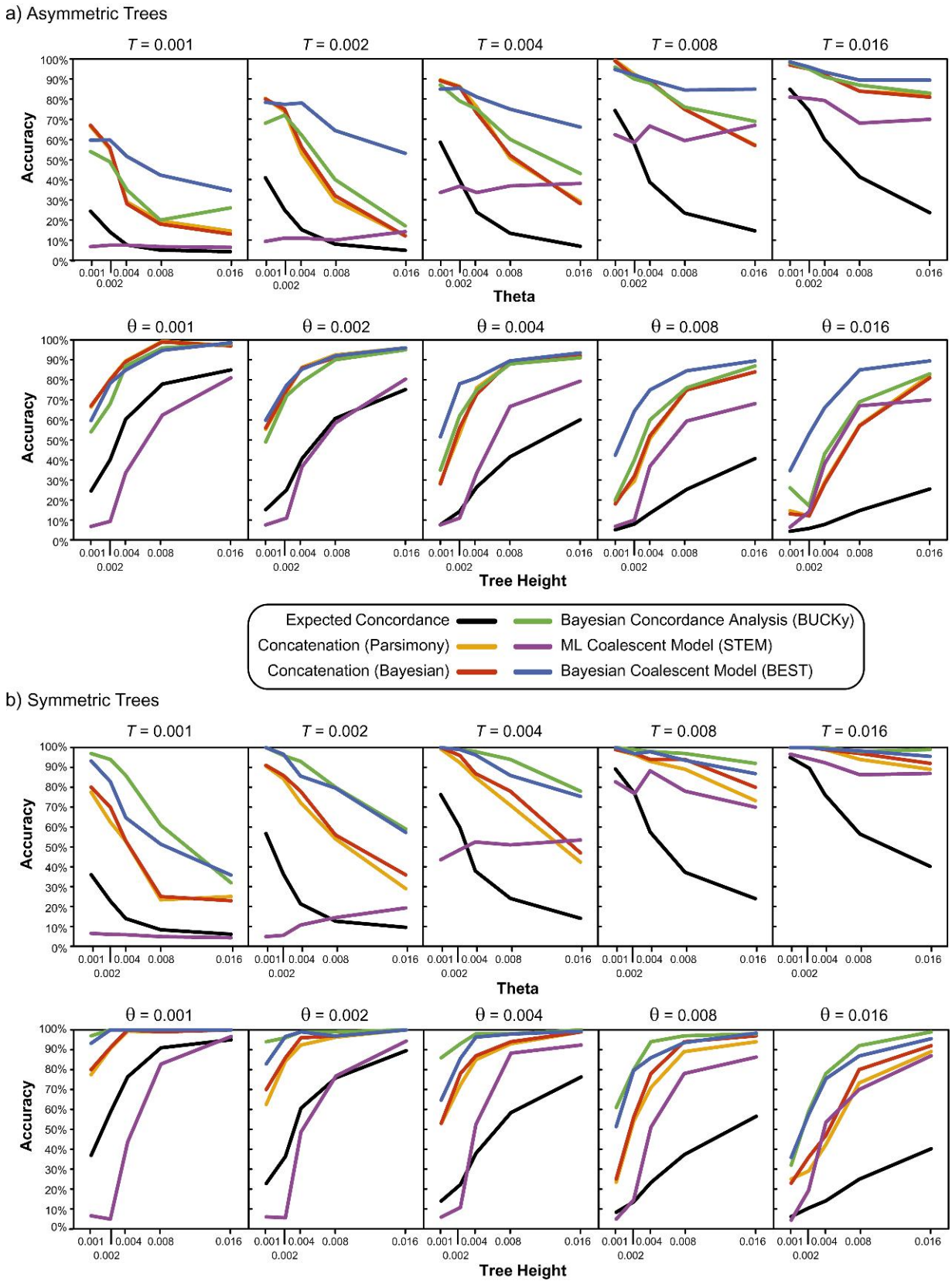


FIGURE 4. Comparison of the accuracy of species tree inference methods under simulation for the (i) asymmetric and the (ii) symmetric species trees using 100 loci. Species tree accuracy is plotted on the *y*-axis in relation to changes in theta (θ) and tree height (*T*).

TABLE 2. Summary of the STEM results from the asymmetric species tree analyses using 100 loci

| $T$ | θ | Average number of tied ML trees | Percentage of times the correct species tree is in the set of tied ML trees (%) | Percentage of replicates producing one ML tree (%) | Accuracy of STEM when producing one ML tree | |
|---|---|---|---|---|---|---|
| | | | | | Rooted (%) | Unrooted (%) |
| 0.001 | 0.001 | 14.52 | 100 | 0 | — | — |
| 0.001 | 0.002 | 13.20 | 98 | 0 | — | — |
| 0.001 | 0.004 | 12.78 | 93 | 0 | — | — |
| 0.001 | 0.008 | 20.28 | 81 | 0 | — | — |
| 0.001 | 0.016 | 61.68 | 79 | 0 | — | — |
| 0.002 | 0.001 | 14.16 | 96 | 0 | — | — |
| 0.002 | 0.002 | 17.06 | 86 | 2 | 50.0 | 50.0 |
| 0.002 | 0.004 | 23.76 | 72 | 3 | 0.0 | 33.3 |
| 0.002 | 0.008 | 26.28 | 73 | 6 | 16.7 | 50.0 |
| 0.002 | 0.016 | 18.04 | 53 | 16 | 6.3 | 25.0 |
| 0.004 | 0.001 | 3.50 | 60 | 41 | 31.7 | 63.4 |
| 0.004 | 0.002 | 3.46 | 54 | 49 | 22.4 | 53.1 |
| 0.004 | 0.004 | 2.76 | 60 | 48 | 39.6 | 52.1 |
| 0.004 | 0.008 | 3.28 | 49 | 58 | 22.4 | 43.1 |
| 0.004 | 0.016 | 2.44 | 44 | 70 | 34.3 | 40.0 |
| 0.008 | 0.001 | 1.42 | 61 | 85 | 60.0 | 74.1 |
| 0.008 | 0.002 | 1.48 | 52 | 82 | 46.3 | 67.1 |
| 0.008 | 0.004 | 1.26 | 61 | 87 | 58.6 | 72.4 |
| 0.008 | 0.008 | 1.36 | 56 | 88 | 55.7 | 63.6 |
| 0.008 | 0.016 | 1.12 | 61 | 94 | 58.5 | 67.0 |
| 0.016 | 0.001 | 1.10 | 75 | 95 | 73.7 | 85.3 |
| 0.016 | 0.002 | 1.12 | 73 | 94 | 72.3 | 81.9 |
| 0.016 | 0.004 | 1.06 | 68 | 97 | 67.0 | 81.4 |
| 0.016 | 0.008 | 1.02 | 58 | 99 | 57.6 | 66.7 |
| 0.016 | 0.016 | 1.02 | 61 | 99 | 60.6 | 67.7 |

Notes: The percentage of STEM analyses producing a single ML tree increases with T, as does the accuracy of these ML point estimates for both rooted and unrooted trees. For any specific value of T, there is a greater chance of finding the correct species tree in the set of tied ML trees when θ is low.

incorrect species tree (Table 2). However, recalculating the accuracy of STEM using only the most informative results (i.e., using only those analyses that resulted in one ML tree) does not result in an significant increase in accuracy (Table 2).

### *Credibility Intervals and Posterior Probability Accuracy*

The average number of species trees contained in the 95% credible intervals from the BEST analyses decreases steadily with increasing tree heights, and when $T=0.016$ the average number of trees in the 95% credible interval is $< 2$ (Table 3). When $T = 0.001$, the 95% credible intervals contain anywhere from 1 to 76 species trees (Table 3). This variability in the size of the credible set likely reflects the variability in waiting times between speciation events in the simulated species trees. Increasing the parameter θ tends to result in a larger number of species trees in the 95% credible intervals, although this relationship is not strict and intermediate values of θ often have the smallest credible sets (Table 3).

TABLE 3. Summary statistics of the 95% credible intervals of trees obtained from the BEST analyses

| | | Population size (θ) | | | | |
|---|---|---|---|---|---|---|
| | | 0.001 | 0.002 | 0.004 | 0.008 | 0.016 |
| Tree height ($T$) | 0.001 | 90% (9.89: 1–76) | 95% (8.35: 2–52) | 95% (9.15: 2–64) | 93% (11.37: 3–58) | 83% (15.22: 2–73) |
| | | 99% (5.36: 1–22) | 98% (5.07: 1–25) | 97% (5.82: 1–19) | 97% (10.52: 1–66) | 87% (13.82: 1–58) |
| | 0.002 | 91% (3.97: 1–13) | 99% (3.65: 1–14) | 97% (4.33: 1–40) | 96% (4.04: 1–14) | 94% (6.27: 1–50) |
| | | 100% (2.73: 1–10) | 100% (2.75: 1–12) | 100% (3.17: 1–22) | 97% (3.64: 1–16) | 96% (5.81: 1–47) |
| | 0.004 | 99% (2.85: 1–13) | 98% (2.77: 1–13) | 96% (2.5: 1–12) | 96% (2.85: 1–12) | 95% (4.03: 1–40) |
| | | 96% (2.13: 1–5) | 100% (1.99: 1–6) | 100% (1.79: 1–4) | 98% (2.04: 1–13) | 94% (2.8: 1–12) |
| | 0.008 | 100% (1.93: 1–12) | 100% (1.9: 1–11) | 100% (1.88: 1–8) | 98% (1.95: 1–8) | 98% (2.29: 1–9) |
| | | 98% (1.77: 1–3) | 97% (1.64: 1–3) | 100% (1.49: 1–3) | 99% (1.49: 1–3) | 99% (1.89: 1–6) |
| | 0.016 | 100% (1.43: 1–3) | 100% (1.42: 1–8) | 100% (1.48: 1–6) | 100% (1.53: 1–3) | 99% (1.79: 1–11) |
| | | 95% (1.35: 1–3) | 96% (1.33: 1–3) | 97% (1.31: 1–3) | 100% (1.33: 1–3) | 99% (1.39: 1–3) |

Notes: Results are shown for the asymmetric species trees (above) and maximally symmetric species trees (below) using 100 loci. The coverage probability (the percentage of times the true species tree is contained in the 95% credible interval) tends to increase with increasing tree heights ($T$), while the average number of trees contained in the 95% credible interval (listed in parentheses with the minimum and maximum) decreases.

The coverage probabilities (the percentage of times the true species tree is contained in the 95% credible interval) for the BEST analyses tend to increase with $T$ and decrease with θ (Table 3). In most cases, the coverage is greater than the nominal value of 95%. Under the most difficult demographic scenario ($T = 0.001$ and $\theta = 0.016$; $0.25N_e$), however, the correct species tree is present in only 83% of the credible sets of trees. Coverage probabilities reach 100% when $\theta \leqslant 0.008$ and $T \geqslant 0.002$ (Table 3).

In general, the correlation between phylogenetic accuracy and the posterior probability values obtained from BEST is high, although overestimation does occur under certain demographic scenarios (Fig. 5). For the asymmetric trees, overestimation occurs when $\theta = 0.001$, resulting in posterior probability values of 0.85, for example, having an accuracy of only 0.57 (Fig. 5a). This particular value of θ is the furthest from the prior mean that we used in our analyses (prior mean $\theta = 0.015$). Thus, it appears that a misspecified prior for θ may lead to inflated estimates of posterior probabilities. To compensate for the overestimate of posterior probability support for those clades with high support, the accuracy of clades with low posterior probability (e.g. posterior probability $< 0.5$) is underestimated (Fig. 5a). The largest discrepancy in accuracy for the symmetric trees occurs at midrange posterior probability values (e.g., 0.5–0.7) when $\theta = 0.001$ (Fig. 5b). However, posterior probability values $\geqslant 0.8$ tend to be either accurate or underestimates of the true accuracy, with the exception of a slight overestimation when $\theta = 0.016$ (Fig. 5b).

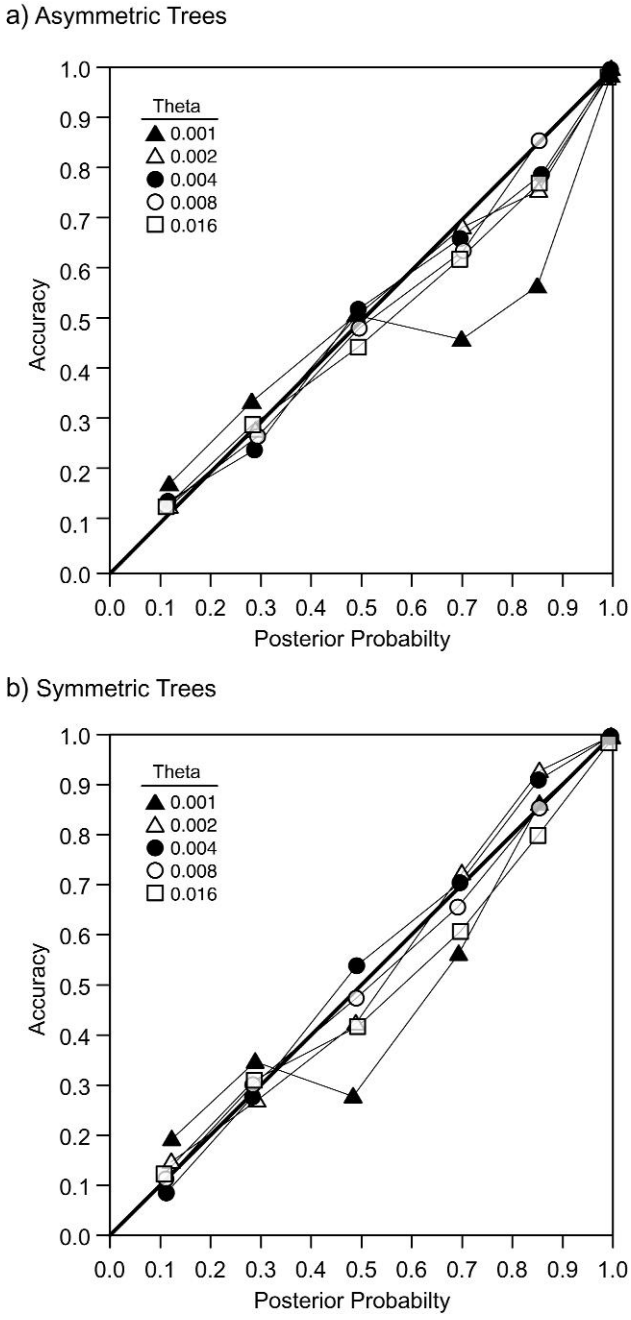


FIGURE 5. Accuracy of posterior probability values obtained from BEST for the (i) asymmetric and the (ii) symmetric species trees for different values of theta (θ). Average posterior probability values are based on the analyses using 100 loci.

## DISCUSSION

### *The Influence of Demographic History on Species Tree Estimation*

The speciation history of a clade has a large impact on the performance of species tree inference methods. In general, the expected time for a large sample of alleles at a neutral locus to find their most recent common ancestor is $4N_e$ generations (Tajima 1983). Our simulation covers a wide range of coalescent units ranging from $0.25N_e$ to $64N_e$, which are likely to describe the speciation histories typical of many empirical studies. For instance, studies of recent radiations often have to contend with large population sizes and recent divergence events (Belfiore et al. 2008), and this challenging evolutionary scenario is reflected by coalescent units $\leqslant 4N_e$.

Within the context of the simulations presented here (e.g., two topologies for five species and exemplar sampling), all methods of analysis appear to be highly accurate when $T$ is large and θ is small (Fig. 3), which represents a "best case" demographic scenario where the probability of deep coalescence is minimized. Increasing the probability of deep coalescence by increasing θ has minimal impacts on the performance of methods when $T$ is large (Fig. 4), and this should be a reassuring result for empiricists studying clades that have a relatively "deep" phylogenetic history. When more challenging scenarios are created by reducing $T$ and increasing the probability of deep coalescence (i.e., coalescent units $< 4N_e$; Fig. 4), all methods of analysis examined here run into trouble. Inferring species trees using BEST (and to some degree BUCKy) outperforms the other methods in these most difficult situations, but the overall accuracy for the estimated species trees remains low (Figs. 3 and 4). Sequencing additional loci

has the potential to greatly improve species tree accuracy (Edwards et al. 2007); however, under the most difficult demographic scenarios examined here, simulations with 100 loci were still insufficient to produce highly accurate results.

### *Species Tree Accuracy using BEST*

BEST appears to overestimate posterior probabilities when the chosen prior for θ assigns a very low probability density to the true parameter value. In conventional Bayesian phylogenetic analysis, it is known that model overspecification does not tend to bias posterior probabilities, while model underspecification leads to posterior probabilities that are too high on average (Huelsenbeck and Rannala 2004). In this context, an overspecified model is not necessarily "incorrect" when the true model is nested within it (by imposing certain fixed relationships among the extraneous parameters), while the underspecified model is strictly incorrect. Thus, it appears that a misspecified prior leads to results similar to those obtained with an incorrect model—posterior probabilities that overestimate confidence. Overall, it is encouraging that in our study posterior probabilities greater than 0.8 tended to be highly accurate (Fig. 5).

To investigate model misspecification of θ further, we reran the BEST analyses of the data simulated with $\theta = 0.001$ (on the asymmetric species tree) with a more reasonable prior mean ($\theta = 0.0015$) than that used in the previous analyses ($\theta = 0.015$). Overestimation of the posterior probability values is less extreme when the prior value for θ is closer to the true parameter value. For example, posterior probability values of 0.85 have an average accuracy of 0.70 (result not shown) versus an accuracy of only 0.57 when the prior is misspecified (Fig. 5). Further simulation studies could be helpful in clarifying whether priors on θ exist that are "conservative" leading to posterior probabilities that are accurate (or perhaps underestimate accuracy) when large.

### *Suggestions for Empirical Studies*

Empiricists should be aware of the benefits and limitations associated with different species tree inference approaches when designing a study and selecting a particular method of analysis. Although a comparison of different approaches is often desirable, many species tree inference methods are not currently capable of handling large multilocus data sets composed of many species, dense population sampling, or large numbers of loci. Many users experience difficulties analyzing large data sets exceeding approximately 50 samples, and the Bayesian methods in particular have trouble reaching stationarity. Some of these difficulties can be overcome by ensuring that the prior for θ is appropriate (Leaché 2009), but this is not a panacea for every MCMC convergence problem. Sampling multiple individuals within species is desirable because it increases species tree accuracy (Maddison and Knowles 2006; Liu et al. 2008; Heled and Drummond 2010); however, including "too many" samples may hinder the convergence of MCMC analyses. Summary statistic approaches for estimated species trees may be desirable under these circumstances because they can accommodate larger numbers of samples (Liu et al. 2009). A potential drawback of some species tree inference methods is the assumption that species assignments are known with certainty prior to the analysis. This assumption is required for the multispecies coalescent methods used here (BEST and STEM). This assumption can be thought of as a trade-off because establishing species assignments a priori enables these methods to benefit from the multispecies coalescent model (Rannala and Yang 2003; Liu et al. 2009). Species assignments must be made carefully because assignment errors can mislead species tree inference and result in strong support for an incorrect species tree (Leaché 2009). Obtaining a multilocus phylogenetic estimate that is assumption-free regarding species membership has certain advantages for studies of recently diverged taxa and species complexes at the phylogeographic level, including the opportunity to discover new lineages, test the exclusivity of populations and species, and track down specimens crossing species or population boundaries. New species tree inference methods that attempt to treat the number of species, and the assignment of individuals to species, as unknown variables are a step in the right direction (O'Meara 2010).

The results of our simulation study suggest that most methods of analysis outperform STEM under a wide range of conditions. Previous simulation studies have suggested that the accuracy of STEM is higher compared to data concatenation (Kubatko et al. 2009) and to the minimize deep coalescences method (McCormack et al. 2009). The accuracy of STEM should be high when gene trees are known with certainty, but our simulation strategy produced some challenging data sets containing very little information from which to infer accurate gene trees. As the information content available in the simulated data diminishes, STEM results in an increased number of species trees that are tied for the ML score (Table 2). STEM is not misleading under these circumstances, rather, STEM is uninformative. Users of STEM should always examine a collection of trees in the neighborhood of the ML tree to gain a sense of the informativeness of the analysis, especially in cases where the accuracy of the estimated gene trees (and branch lengths) is suspected to be low. It is common for empirical data to produce genealogies containing ambiguities in relationships and branch length uncertainty, and from our perspective, species tree inference methods that can accommodate this uncertainty have an advantage over methods that rely on point estimates.

### *Future Directions*

There are several aspects of our simulation strategy that can be expanded upon to capture additional

complexities inherent to real empirical data sets. First, future simulation studies should consider more complex evolutionary histories involving more species, additional substitution model complexity, and population substructure. Second, the impacts of missing data on species tree inference remain unstudied, and this seems like an important area of investigation because it is often difficult to collect multilocus data sets with complete representation for each sample. Finally, an underlying assumption of species tree inference is that gene tree discordance is the result of deep coalescence, and it is unclear how the processes of gene flow, recombination, selection, and gene duplication and extinction will impact species tree accuracy.

## Supplemental Material

Supplementary material can be found at http://www.sysbio.oxfordjournals.org/.

## Funding

This work was supported by a National Science Foundation Postdoctoral Research Fellowship in Biological Informatics (DBI-0805455) awarded to A.D.L.

## Acknowledgements

We thank M. Fujita, J. Inoue, Y. Wang, and Z. Yang for help with various aspects of the research. We also thank F. Burbrink, L. Knowles, L. Kubatko, L. Liu, B. Moore, three anonymous reviewers, and the Huelsenbeck, Moritz, and McGuire laboratory at UC Berkeley for their insightful comments and suggestions.

## References

Ané C., Larget B., Baum D.A., Smith S.D., Rokas A. 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24:412–426.

Baum D.A. 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. Taxon. 56:417–426.

Belfiore N.M., Liu L., Moritz C. 2008. Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae). Syst. Biol. 57:294–310.

Burgess R., Yang Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing error. Mol. Biol. Evol. 25: 1979–1994.

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genetics. 2:762–768.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference, and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Eckert, A.J., Carstens B.C. 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Mol. Phylogenet. Evol. 49:832–842.

Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63(1):1–19.

Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. U. S. A. 104:5936–5941.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.

Hey J., Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics. 167:747–760.

Hillis D.M. 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. 44:2–16.

Hillis D.M., Huelsenbeck J.P., Swofford D.L. 1994. Hobgoblin of phylogenetics. Nature. 369:363–364.

Huelsenbeck J.P., Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53: 904–913.

Huang H., Knowles L.L. 2009. What is the danger of the anomaly zone for empirical phylogenetics? Syst. Biol. 58:527–536.

Jennings W.B., Edwards S.V. 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. Evolution. 59:2033–2047.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

Kingman J.F.C. 1982. On the genealogy of large populations. Stoch. Proc. Appl. 13:235–248.

Kingman J.F.C. 2000. Origins of the coalescent: 1974–1982. Genetics. 156:1461–1463.

Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics. 25:971–973.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56: 17–24.

Leaché A.D. 2009. Species tree discordance traces to phylogeographic clade boundaries in North American fence lizards (*Sceloporus*). Syst. Biol. 58:547–559.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics. 24:2542–2543.

Liu L., Pearl D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

Liu L., Pearl D.K., Brumfield R.T., Edwards S.V. 2008. Estimating species trees using multiple-allele DNA sequence data. Evolution. 62:2080–2091.

Liu L., Yu L., Kubatko L., Pearl D.K., Edwards S.V. 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320–328.

Liu L., Yu L., Pearl D.K., Edwards S.V. 2009. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58: 468–477.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46: 523–536.

Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

McCormack J.E., Huang H., Knowles L.L. 2009. Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. Syst. Biol. 58:501–508.

Nylander J.A.A., Wilgenbusch J.C., Warren D.L., Swofford D.L. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics. 24(4):581–583.

O'Meara B.C. 2010. New heuristic methods for joint species delimitation and species tree inference. Syst. Biol. 59:59–73.

Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

Rannala B., 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., Yang Z. 2008. Phylogenetic inference using whole genomes. Annu. Rev. Genomics Hum. Genet. 9:217–231.

Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798–804.

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

Rosenberg N.A., Tao R. 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. Syst. Biol. 57: 131–140.

Swofford D L. 2001. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Tajima R. 1983. Evolutionary relationships of DNA sequences in finite populations. Genetics. 105:437–460.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 27:1586–1591.

Yang Z. 2009. MCMCcoal version 1.2. Available from: http://abacus.gene.ucl.ac.uk/software/MCMCcoal.html.

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