Robustness to Divergence Time Underestimation When Inferring Species Trees from Estimated Gene Trees

MICHAEL DEGIORGIO1, 2 AND JAMES H. DEGNAN3, ∗

1Department of Integrative Biology, University of California, Berkeley, CA 94720, USA; 2Department of Biology, Pennsylvania State University, University Park, PA 16802, USA; and 3Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131, USA

∗Correspondence to be sent to: Department of Mathematics and Statistics, University of New Mexico, 1 University of New Mexico, Albuquerque, NM 87131, USA; E-mail: jamdeg@gmail.com.

Received 31 March 2013; revised returned 7 June 2013; accepted 20 August 2013

Associate Editor: Robb Brumfield

Abstract.—To infer species trees from gene trees estimated from phylogenomic data sets, tractable methods are needed that can handle dozens to hundreds of loci. We examine several computationally efficient approaches—MP-EST, STAR, STEAC, STELLS, and STEM—for inferring species trees from gene trees estimated using maximum likelihood (ML) and Bayesian approaches. Among the methods examined, we found that topology-based methods often performed better using ML gene trees and methods employing coalescent times typically performed better using Bayesian gene trees, with MP-EST, STAR, STEAC, and STELLS outperforming STEM under most conditions. We examine why the STEM tree (also called GLASS or Maximum Tree) is less accurate on estimated gene trees by comparing estimated and true coalescence times, performing species tree inference using simulations, and analyzing a large data set keeping track of false positive and false negative rates for inferred clades. We find that although true coalescence times are more ancient than speciation times under the multispecies coalescent model, estimated coalescence times are often more recent than speciation times. This underestimation can lead to increased bias and lack of resolution with increased sampling (either alleles or loci) when gene trees are estimated with ML. The problem appears to be less severe using Bayesian gene-tree estimates. Bayesian, false negative; false positive; gene tree; maximum likelihood; phylogenetics; Robinson-Foulds; statistical consistency.

Two of the most widely used programs for inferring species trees from multilocus data are BEST (Liu and Pearl 2007) and *BEAST (Heled and Drummond 2010). Although these methods tend to have the best performance for moderately sized data sets, they are computationally intensive and require estimating a joint posterior distribution for gene trees and a species tree. This posterior distribution has an increasing number of parameters as the number of loci increases, making it difficult for the programs to achieve convergence for data sets with many loci (Cranston et al. 2009; Kubatko et al. 2011). This limitation motivates using computationally efficient methods to handle large phylogenomic data sets. Recent examples include using the Maximum Pseudolikelihood Estimate of the Species Tree (MP-EST; Liu et al. 2010a) and the Species Tree estimation using Average Ranks (STAR; Liu et al. 2009) methods to estimate a mammalian phylogeny from 447 loci (Song et al. 2012), and STAR to estimate a bird phylogeny from 1541 loci (McCormack et al. 2013a).

Methods for inferring species trees from gene trees can be partitioned into those that use branch lengths (or coalescence times) in the gene trees and topology-based methods. Topology-based methods include Species Tree inference with Likelihood for Lineage Sorting (STELLS), which maximizes the probability of the observed gene-tree topologies (Wu 2012), maximizing the pseudo-likelihood of all rooted triples in observed gene trees (MP-EST; Liu et al. 2010a), the STAR method (Liu et al. 2009), and the minimize deep coalescence approach (Maddison and Knowles 2006; Than and Nakhleh 2009). Methods using coalescence times include Species Tree Estimation using Average Coalescence times (STEAC, Liu et al. 2009) and Global LATeSt Split (GLASS) / Maximum Tree (Liu et al. 2010b; Mossel and Roch 2010), as well as variations on these methods (Jewett and Rosenberg 2012; Helmkamp et al. 2012). Although we examine most of these methods in this article, our focus is on understanding why the GLASS tree has not performed as well as most other methods on gene trees estimated from DNA sequences.

STEAC, STAR, and GLASS/Maximum Tree are all similar in that they first construct a distance matrix, where each entry in the matrix is an estimated distance between a pair of taxa. For STEAC, the distance between a pair of taxa is entered as twice the average estimated coalescence time observed on the estimated gene trees, whereas GLASS uses the minimum estimated pairwise coalescence time. The distance matrix used by STAR is similar to that of STEAC, but is obtained by setting all internal branch lengths of the input gene trees to 1.0 units, and making the external branches have a minimum length of 1.0 with lengths added to external branches to make the tree ultrametric (i.e., a constant distance from tips of a tree to the root). This makes the STAR tree only dependent on the topologies of the gene trees. For STEAC and STAR, neighbor joining or UPGMA is typically applied to the distance matrix, whereas for GLASS, single-linkage clustering is typically applied, where the distance from one set of taxa to another set is taken to be the minimum of all pairwise distances rather than the average (the approach taken by UPGMA).

Species tree inference using minimum interspecific coalescence times (where the minimum is also taken over multiple loci) was proposed independently by Liu et al. (2010b) and Mossel and Roch (2010). Liu et al.
2014 DEGIORGIO AND DEGNAN—INFERRING SPECIES TREES FROM ESTIMATED GENE TREES 67

trees (Roch 2013). Proven to be more efficient than STEAC and STAR when (Rannala and Yang 2003). The GLASS method can be species tree under the multispecies coalescence model is also a maximum likelihood estimate (MLE) of the times. Mossel and Roch (2010) showed that the GLASS tree is a statistically consistent estimator of the species tree under some conditions, whereas Liu et al. (2010b) showed that if population sizes are constant throughout the species tree and gene trees are known with certainty, then the GLASS tree constructed from known distances is also a maximum likelihood estimate (MLE) of the species tree under the multispecies coalescence model (Rannala and Yang 2003). The GLASS method can be proven to be more efficient than STEAC and STAR when inferring three-taxon species trees from known gene trees (Roch 2010).

The GLASS estimator has been implemented in several software packages, including STEM (Kubatko et al. 2009), Phybase (Liu and Yu 2010), and PhyloNet (Thompson et al. 2008). We emphasize that these programs all perform functions other than simply estimating the GLASS tree. For example, STEM returns likelihoods for species trees near the MLE or for user-specified species trees, whereas Phybase and PhyloNet implement several other methods for estimating species trees and species networks (Yu et al. 2011, 2012).

In addition to the properties of statistical consistency when gene trees are known with certainty and being the MLE of the species tree, Mossel and Roch (2010) provided a condition under which the GLASS tree is statistically consistent when gene trees are estimated with error. Unfortunately, simulations have shown that although performance can be good when gene trees are known, the GLASS estimate of the species tree can become worse as the number of estimated gene trees increases (Leaché and Rannala 2011; Wu 2012).

The disconnect between the apparently favorable theoretical properties reported for GLASS and its poor performance on gene trees estimated from simulated DNA sequences illustrates the importance of considering the sensitivity of species tree inference methods to noise introduced by mutation (Huang and Knowles 2009; Huang et al. 2010). We show that the GLASS method is statistically inconsistent when applied to gene trees estimated with error using maximum likelihood (ML) under reasonable substitution models. In particular, as the number of loci gets large, the conditions assumed for statistical consistency in Mossel and Roch (2010) do not apply under standard substitution models, and all minimum estimated pairwise coalescence times eventually become zero. This situation leads to methods returning either a completely unresolved tree, which is what frequently occurs in the STEM and Maximum Tree implementations, or an arbitrarily chosen tree.

To illustrate some of the difficulties encountered by GLASS and related methods, we perform two simulation studies as well as a study using empirical data. In the first simulation study, we consider a fixed species-tree topology and examine the coalescence times estimated on the individual gene trees using ML under different parameters (differing alignment lengths, two branch-length settings, and two values of \( \theta \), the population-scaled mutation rate). This scenario was used to examine the proportion of gene trees that would be estimated accurately enough for the conditions provided by Mossel and Roch (2010) that lead to consistency of GLASS under error in the gene trees. We also examine differences between estimated and true coalescence times, as well as the bias in species divergence times when STEM versions 1.1 and 2.0 are used to estimate the species tree. The second simulation study investigates the accuracy of MP-EST, STAR, STEAC, STELLS, and STEM versions 1.1 and 2.0 averaged over species trees as a function of the number of loci and compares the results of using ML versus Bayesian methods to estimate gene trees. The empirical study also evaluates the accuracy of STEM and other methods as a function of the number of loci for ML versus Bayesian estimated gene trees, using a subset of the great ape data from Burgess and Yang (2008).

We note that there are important differences between STEM1.1 and STEM2.0 in terms of how external branches of length zero are handled. In particular, STEM2.0 finds the minimum horizontal distance between each pair of taxa, uses these nonzero distances to construct a distance matrix, and constructs an estimated species tree from this distance matrix. (STEM2.0 only reports a distance of zero for a pair of taxa if all loci have a distance of 0.) STEM1.1, on the other hand, uses branch lengths of zero in the distance matrix (although it adds a small increment to make the distance nonzero). As an example, consider the pair of input gene trees

\[
(A:0.01,B:0.01):0.03,C:0.04);
\]

\[
((A:0.00,C:0.00):0.02,B:0.01);
\]

Using \( \theta = 0.01 \) in the settings files, STEM1.1 returns

\[
(B:1.000000,(A:0.000001,C:0.000001):0.999999);
\]

whereas STEM2.0 returns

\[
(C:2.000000,(A:1.000000,B:1.000000):1.000000);
\]

Here the values in the estimated species trees are the minimum coalescence times divided by the value of \( \theta \) in the settings file to convert to units proportional to coalescent units. STEM2.0 does not use the divergence time of 0.0 for \( A \) and \( C \) from the second locus, but does use the divergence time of 0.02 between \( B \) and \( C \) from the second locus. STEM1.1, however, uses the time between \( A \) and \( C \) from the second gene tree (and adds 10^{-9}). Both
versions of STEM also use the divergence time of 0.01 between A and B from the first locus.

Trees returned by STEM2.0 tend to be at least as resolved as trees returned by STEM1.1; however, this is not always the case. An example is from the two input gene trees

\[
\{(A:0.02, B:0.02) : 0.03, C:0.05 \};
\]

\[
\{(A:0.00, C:0.00) : 0.01, B:0.01 \};
\]

for which the STEM1.1 tree is the same as the previous example, but where the STEM2.0 tree is the unresolved tree

\[
\{C:1.0000, (A:1.00000, B:1.00000):0.00000 \};
\]

In this case, when STEM2.0 computes the distance matrix between the species, the distance between A and C is 0.0, and the minimum nonzero distance between A and C. From the pairwise distances, either A or B and C can be clustered first (since their minimum distances are both 1.0). If A and B are clustered first, then the distance between C and (A, B) is the minimum of the distances between A and B and C. If C is first, then the distance between A and B is 0.01. If 0.001 is used, the STEM2.0 tree is the unresolved tree.

The two methods are identical as long as all pairwise distances for all loci are nonzero (which is the case for known gene trees), but they can differ dramatically both in terms of the estimated topology and the calculated likelihood scores when zero-length branches occur in the gene trees. Although the difference between STEM1.1 and STEM2.0 was initially due to a programming error in STEM2.0, it has been retained in STEM2.0 and STEM-Ly (which allows some degree of hybridization between species) with no plans of changing back to the STEM1.1 model with transition/transversion (ti/tv) ratio of 4.6 and nucleotide frequencies of 0.3, 0.2, 0.2, and 0.3 for A, C, G, and T, respectively. (The ti/tv and nucleotide frequency values are motivated by the great ape data described below.) Gene trees were estimated in PHYLIP (Felsenstein 1989) using ML with a molecular clock using dnamlk under the correct substitution model with estimated base frequencies. Global rearrangements were used with 10 random input orders for the sequences for the ML search. For each setting of N, t, and sequence length, 11,000 independent gene trees were simulated from the species tree.

The maximum error in coalescent times was simulated by comparing estimated and true coalescent times, \( D_i^k - D_j^k \), for each locus k and each pair of taxa \( i \neq j \) in the gene trees. Plots for the distribution of maximum error in estimated coalescence times, \( \max_{ij \in [a,b,c,d]} D_i^k - D_j^k \), and for the estimated and true distributions of coalescence times, \( D_{ac} \) and \( D_{ac} \), respectively, for lineages from species A and C, were generated using a modification of the vioplot package (Hintze and Nelson 1998) in R (R Development Core Team 2012), which generates boxplot-like representations of data called “violin plots” using nonparametric density estimates instead of the traditional five-number summary in a boxplot. Violin plots retain the median (represented by a white dot) and interquartile range (heavy black line segment) information typically reported in boxplots. The package was modified to use the default density function in R for the nonparametric density estimate, and to make side-by-side comparisons of two different densities, similar to side-by-side stem-and-leaf plots. Densities are not drawn with a normalized area, but are drawn so that the maximum height of the density curve is constant. Note also that the true distribution of true coalescence times is shifted exponential, and here looks slightly different due to the nonparametric density estimate. The bias in the estimated speciation time separating species A and C only is reported based on estimating the species tree from STEM1.1 and STEM2.0 using 100 replicates of either 10 loci or 100 loci, resulting in 11,000 gene trees per set. The same gene trees were used as input for STEM1.1 and STEM2.0.

METHODS

Estimated Coalescence Times and Bias Simulations

A fixed species tree with topology \((A, B, C, D)\) and all branch lengths equal to 0 coalescent units, with \( x = 0.5 \) and \( x = 5.0 \) coalescent units (total tree height is either 1N or 10N), was used to simulate the error and bias in estimated coalescence times. Here 1.0 coalescent units equals \( N \) generations, where \( N \) is the effective number of gene copies in a population. Values of \( t \) used were 0.01 and 0.001. Gene trees were simulated in COAL.

Sequences of length 500, 1000, 2000, 4000, and 8000 nucleotides (nt) were simulated in the program SeqGen (Rambaut and Grassly 1997) from each gene tree branch lengths multiplied by \( 0.01 \) under the HKY model with transition/transversion (ti/tv) ratio of 4.6 and nucleotide frequencies of 0.3, 0.2, 0.2, and 0.3 for A, C, G, and T, respectively. (The ti/tv and nucleotide frequency values are motivated by the great ape data described below.) Gene trees were estimated in PHYLIP (Felsenstein 1989) using ML with a molecular clock using dnamlk under the correct substitution model with estimated base frequencies. Global rearrangements were used with 10 random input orders for the sequences for the ML search. For each setting of N, t, and sequence length, 11,000 independent gene trees were simulated from the species tree.

The maximum error in coalescent times was simulated by comparing estimated and true coalescent times, \( D_i^k - D_j^k \), for each locus k and each pair of taxa \( i \neq j \) in the gene trees. Plots for the distribution of maximum error in estimated coalescence times, \( \max_{ij \in [a,b,c,d]} D_i^k - D_j^k \), and for the estimated and true distributions of coalescence times, \( D_{ac} \) and \( D_{ac} \), respectively, for lineages from species A and C, were generated using a modification of the vioplot package (Hintze and Nelson 1998) in R (R Development Core Team 2012), which generates boxplot-like representations of data called “violin plots” using nonparametric density estimates instead of the traditional five-number summary in a boxplot. Violin plots retain the median (represented by a white dot) and interquartile range (heavy black line segment) information typically reported in boxplots. The package was modified to use the default density function in R for the nonparametric density estimate, and to make side-by-side comparisons of two different densities, similar to side-by-side stem-and-leaf plots. Densities are not drawn with a normalized area, but are drawn so that the maximum height of the density curve is constant. Note also that the true distribution of true coalescence times is shifted exponential, and here looks slightly different due to the nonparametric density estimate. The bias in the estimated speciation time separating species A and C only is reported based on estimating the species tree from STEM1.1 and STEM2.0 using 100 replicates of either 10 loci or 100 loci, resulting in 11,000 gene trees per set. The same gene trees were used as input for STEM1.1 and STEM2.0.

Estimating Species Trees Simulations

Species trees were estimated using MP-EST, STAR, STEAC, STELLS, STEM1.1, and STEM2.0 using four species with either one or three individuals sampled per species. Species trees were simulated under a pure birth Yule model using sim.bd.taxa.age in the TreeSim R package (Stadler 2011). The height of the tree (at the most recent common ancestor [MRCA]) was constrained to 1.0 coalescent units and the birth rate was set to

\[
N = 1.00000, (A:1.00000, B:1.00000:0.00000\};
\]

which is \( N = 0.01 \) and \( N = 0.001 \) coalescent units (total tree height is either 1N or 10N), was used to simulate the error and bias in estimated coalescence times. Here 1.0 coalescent units equals \( N \) generations, where \( N \) is the effective number of gene copies in a population. Values of \( t \) used were 0.01 and 0.001. Gene trees were simulated in COAL.

Sequences of length 500, 1000, 2000, 4000, and 8000 nucleotides (nt) were simulated in the program SeqGen (Rambaut and Grassly 1997) from each gene tree branch lengths multiplied by \( 0.01 \) under the HKY model with transition/transversion (ti/tv) ratio of 4.6 and nucleotide frequencies of 0.3, 0.2, 0.2, and 0.3 for A, C, G, and T, respectively. (The ti/tv and nucleotide frequency values are motivated by the great ape data described below.) Gene trees were estimated in PHYLIP (Felsenstein 1989) using ML with a molecular clock using dnamlk under the correct substitution model with estimated base frequencies. Global rearrangements were used with 10 random input orders for the sequences for the ML search. For each setting of N, t, and sequence length, 11,000 independent gene trees were simulated from the species tree.

The maximum error in coalescent times was simulated by comparing estimated and true coalescent times, \( D_i^k - D_j^k \), for each locus k and each pair of taxa \( i \neq j \) in the gene trees. Plots for the distribution of maximum error in estimated coalescence times, \( \max_{ij \in [a,b,c,d]} D_i^k - D_j^k \), and for the estimated and true distributions of coalescence times, \( D_{ac} \) and \( D_{ac} \), respectively, for lineages from species A and C, were generated using a modification of the vioplot package (Hintze and Nelson 1998) in R (R Development Core Team 2012), which generates boxplot-like representations of data called “violin plots” using nonparametric density estimates instead of the traditional five-number summary in a boxplot. Violin plots retain the median (represented by a white dot) and interquartile range (heavy black line segment) information typically reported in boxplots. The package was modified to use the default density function in R for the nonparametric density estimate, and to make side-by-side comparisons of two different densities, similar to side-by-side stem-and-leaf plots. Densities are not drawn with a normalized area, but are drawn so that the maximum height of the density curve is constant. Note also that the true distribution of true coalescence times is shifted exponential, and here looks slightly different due to the nonparametric density estimate. The bias in the estimated speciation time separating species A and C only is reported based on estimating the species tree from STEM1.1 and STEM2.0 using 100 replicates of either 10 loci or 100 loci, resulting in 11,000 gene trees per set. The same gene trees were used as input for STEM1.1 and STEM2.0.
Gene trees were estimated using COAL (Degnan and Salter 2005), and sequence data for each gene tree was simulated using Seq-Gen under the HKY model with ti/tv ratio of 4.6 and base frequencies of 0.3, 0.2, 0.2, and 0.3 for bases A, C, G, and T. Branch lengths simulated in COAL were multiplied by 0.2 in Seq-Gen, where θ = 0.01 and 0.001 for ML and MrBayes analyses (Ronquist and Huelsenbeck 2003). Only θ = 0.01 was used for the simulations with three individuals sampled per species. The number of loci used to estimate the species tree was 10, 20, 30, 40, 50, 60, 80, 100, 300, and 500. Three alignment-length settings were used: empirical, 550 nt at each locus, and 1000 nt at each locus. The empirical distribution for alignment lengths was based on an 8881-locus great ape data set (described below in the Empirical Data Analysis subsection). For each locus, the alignment length was sampled with replacement from the distribution of alignment lengths in the data set, accepting only alignments at least 300 nt long. The mean length of alignments from the great ape data set, conditional on alignments with lengths of at least 300 nt, was 550 nt, with standard deviation of 113 nt. The empirical distribution of alignment lengths appears to be typical of alignment lengths for multilocus data sets used to infer species trees (Supplementary Material, doi:10.5061/dryad.65tn4).

Gene trees were estimated using ML with *dnamlk* in PHYLIP (Felsenstein 1989) and using a Bayesian method with MrBayes version 3.2 (Ronquist and Huelsenbeck 2003), with only ML utilized for scenarios with three individuals per species. Gene trees were estimated under the correct mutation model using an enforced clock with default prior, fixed ti/tv ratio, and base frequencies estimated from the data. MrBayes runs used four chains with 5 × 10^6 generations and convergence was checked by noting that the standard deviation of split frequencies never exceeded 0.01. For each locus, the majority-rule consensus tree reported by MrBayes was used as the estimated gene tree at that locus, a procedure also employed by Huang et al. (2010). Although gene trees sampled in the Markov chain Monte Carlo (MCMC) simulation are clock-like (due to the enforced clock), the consensus tree reported typically deviates slightly from a clock-like tree. Consequently, we used the mean path length method implemented in PATHd8 (Britton et al. 2007) to convert the consensus into clock-like trees for input into STEM1.1 and STEM2.0. Unresolved consensus gene trees were randomly resolved using the *multi2di* function in the *ape* package (Paradis et al. 2004) in R by adding internal branches of length zero. In all cases, identical input gene trees were used for STEM1.1 and STEM2.0 for ML. Bayesian gene trees were analyzed only using STEM1.1 since results were the same for STEM2.0 on Bayesian gene trees due to the absence of divergence times of 0.

For ML estimated gene trees, there were 1000 independent species trees simulated for each combination of alignment length, θ, number of loci, and number of alleles per species. For gene trees estimated using MrBayes, there were 100 independent species trees generated for each combination of parameters. The accuracy of estimated species trees for all methods is recorded using the Robinson–Foulds (RF) distance (Robinson and Foulds 1981), computed as the sum of the number of false positive and false negative clades on an estimated species tree. A clade is a false positive if it occurs on the estimated species tree but not the true species tree, and a false negative if it occurs on the true species tree but not the estimated species tree. We note that returning a completely unresolved tree results in a RF distance of 50% of the maximum when the RF distance is computed as the sum of false negative and false positive rates, and that this is likely to result in lower (lower) RF distances than using randomly resolved trees, which might occur if STEM1.1 output is used without considering arbitrary resolutions to be unresolved.

The accuracy of the estimated species trees from STEM1.1 and STEM2.0 was additionally measured using the number of false positive, false negative, and true positive clades (i.e., clades that are on both the estimated and true species tree). The number of clades on the estimated species tree is the sum of the number of true positive and false positive clades and measures how resolved the estimated species tree is.

In STEM1.1 analyses, a clade was considered to exist in the STEM tree if the length of the branch directly ancestral to the MRCA for all taxa in the clade was longer than 2 × 10^−6. Our choice of this criterion was based on the observation that input gene trees that should have yielded unresolved clades based on distances being zero were sometimes assigned additional branch lengths by STEM of up to 2 × 10^−4, and no branch lengths smaller than 10^−4 and greater than 2 × 10^−5 were observed in any STEM1.1 analysis. We note that cases in which such small branch lengths are observed also correspond to cases in which the likelihood score returned by STEM1.1 is tied or nearly tied for different resolutions of the clade. For STEM2.0, which does not add lengths of 10^−6 to branches of length zero, we considered a clade to exist in the STEM tree as long as the branch length above the MRCA for all taxa in the clade was strictly positive. Species trees for STEAC and STAR were estimated using UPGMA applied to their distance matrices, so that an outgroup was not needed to root the trees.

**Empirical Data Analysis**

A subset of the data provided in Burgess and Yang (2008) was analyzed. We started with the *neutralA* data set from Burgess and Yang (2008), which contains 14,663 alignments from intragenic regions of autosomal chromosomes, of which 9861 alignments have no missing taxa. These alignments include one sequence each from human, chimpanzee, gorilla, orangutan, and macaque. A molecular clock was tested using a likelihood ratio test (Felsenstein 1981) under the HKY model using the ti/tv ratio of 4.6 estimated in Burgess and Yang (2008) and using empirical base
frequencies. There were 8881 alignments used that passed the molecular clock at the α = 0.05 level. We refer to alignments based on the order in which they occur in the neutral/Δ data set of 14,663 alignments.

ML gene trees were estimated in PHYLIP under the HKY model with τ/ν ratio of 4.6 and estimated base frequencies. Bayesian gene trees were estimated in MrBayes with the same settings as the simulations except that chains were run for 10^6 generations. All species tree methods except STEM2.0 were used on both ML and Bayesian gene trees, with STEM2.0 only being run on ML gene trees. In addition to running STEM on full data sets, we used STEM on random subsets of the alignments using 10, 20, ..., 60, 80, 100, 300, 500, 1000, 2000, and 3000 loci. The subsets were simple random samples chosen without replacement from the set of 8881 molecular clock alignments, and in separate analyses on subsets of alignments at least 300 nt in length (7582 alignments) and at least 600 nt in length (3171 alignments). We used 1000 iterations for each number of loci and each setting. Accuracy of estimated species trees was measured as in the simulation study.

THEORETICAL BACKGROUND

Inconsistency of GLASS on Estimated Gene Trees: Theory

We provide a theoretical argument that, under reasonable assumptions, GLASS is statistically inconsistent when applied to gene trees estimated using ML, in the sense that as the number of loci goes to infinity, the probability that GLASS returns the species-tree topology does not go to 1 when gene trees are estimated using ML.

For notation, we let upper case letters denote species and lower case letters denote genes sampled from corresponding species. We let \( D_{ij}^{(k)} \) be the minimum divergence times for genes sampled from species \( I \) and \( J \) at locus \( k \). Then \( D_{ij} = \min_k D_{ij}^{(k)} \) denotes the minimum coalescence time between species \( I \) and \( J \) across the set of sampled gene trees. The values \( D_{ij} \) are used as a distance matrix, with \( I, J \in \{1, 2, ..., n\} \), and single linkage clustering is applied to the distance matrix to cluster the species.

Despite the fact that GLASS performs well on known gene trees, we show that the conditions referred to by Mossel and Roch (2010), under which GLASS was stated to be statistically consistent in the presence of “moderate” error in the estimated gene trees, do not apply under reasonable assumptions. The sufficient condition for statistical consistency given by Mossel and Roch (2010) can be written as

\[
D_{ij}^{(k)} = \max_{l \neq j} \left| D_{il}^{(k)} - D_{lj}^{(k)} \right| < m/2 \quad \text{for all loci } k.
\]

Condition (1) is motivated by the fact that the minimum true coalescence time \( D_{ij} \) should be close to the species divergence time \( t_{ij} \). Thus, if each estimated coalescence time is close enough to the true coalescence time, then the estimated coalescence time will also be close to the species divergence time. For example, suppose that the species tree has a subtree \( (Ax, Bx)y_C, Cx+y \) and that a single lineage is sampled per species, and suppose that \( y < x \). Recalling that \( m/2 \) is half the length of the shortest branch on the species tree, the condition that each estimated coalescence time differs by less than \( m/2 \) guarantees that \( D_{AC}, D_{BC} > x + y/2 \). With enough loci, it is guaranteed that at least one locus will satisfy \( D_{AB}^{(k)} < x + y/2 \) in which case, the GLASS tree will correctly have the relationships \( D_{AB} < D_{AC}, D_{BC} \).

Although condition (1) would be sufficient for consistency even with estimated gene trees, under typical assumptions, condition (1) fails with probability 1 as the number of loci goes to infinity when distances are estimated using ML using sequences of bounded length. In particular, there is some positive probability that two sequences from different species will be identical in a finite alignment. If the sequences from lineages \( a \) and \( b \) are identical at locus \( k \), then the minimum estimated coalescence time is zero using ML. For any pair of sequences, as the number of loci increases, eventually the sequences will be identical for some locus, and their minimum estimated distance will be zero under ML. Because all minimum estimated distances will eventually be zero, the GLASS tree will be unresolved with all branches having length zero. We make the following assumptions to formalize the argument:

1. If 2 sequences at a locus \( k \) are identical for lineages \( a \) and \( b \), then \( D_{ab}^{(k)} = 0 \).
2. The probability of a mutation on a branch of length \( t > 0 \) on a gene tree is \( p(t) \), with \( 0 < p(t) < 1 \).
3. Along a gene tree, sites are independent (i.e., the probability of a mutation at one site does not depend on the presence of mutation at another site).
4. The number of sites in an alignment is bounded by some finite positive integer \( Y < \infty \).
5. Loci are independent.
6. Lineages sampled from two distinct species \( A \) and \( B \) coalesce at some time \( T \), which is a random
variable with density \( f(t) \) whose support is 0 for \( t < \tau_{AB} \), where \( \tau_{AB} \) is the species divergence time for species \( A \) and \( B \). That is, no gene flow occurs between species \( A \) and \( B \) more recently than the divergence time \( \tau_{AB} \).

**Proposition 1.** Under assumptions 1–6, the probability that condition (1) is satisfied approaches 0 as the number of loci approaches infinity.

A proof of Proposition 1 is provided in Appendix A.

**Zero Divergence Under the \( N_r \) Model**

Although asymptotically (in the number of loci), GLASS does not return a resolved tree when input trees are estimated using ML, the behavior of GLASS on estimated gene trees with finite numbers of loci can be explained, to some extent, through a simple substitution model to analytically investigate the probability that sequences from two species are undiverged at a locus. If two sequences from a pair of nonsister taxa are undiverged at any locus, then GLASS will return either an incorrectly resolved tree or a tree that is not fully resolved. We use probabilities that pairs of sequences are identical to explore the parameter space in which GLASS is especially likely to return unresolved species trees as a function of divergence time, mutation rate, population size, number of loci, and sequence length. For tractability, we use the \( N_r \) substitution model, in which there are \( r > 1 \) character states with equal probabilities of mutating into any other state. The Cavender–Farris–Neyman and Jukes–Cantor models are recovered for \( r = 2 \) and \( r = 4 \), respectively.

The probability that the aligned sequences of length \( y \) for lineages \( a \) and \( b \) are identical at a given locus, averaged over all possible coalescence times, is

\[
P(D_{y}^{AB} = 0) = \sum_{i=0}^{y} \binom{y}{i} \frac{1}{r^i} \left(1 - \frac{1}{r}\right)^{y-i} \frac{e^{-r \mu \tau} (y-i)! \tau_{AB}^i}{2N_\mu \left(\frac{1}{2}\right)^{y-i} + 1}
\]

which is derived in Appendix B.

The parameter \( \theta = 2N_\mu \) occurs in the denominator of equation (2), and the species divergence time measured in mutation units, \( \mu \tau_{AB} \), occurs in the argument of the exponential function. Writing the probability in this form allows considering the effect of changing either \( N \) (which affects \( \theta \) without altering \( \mu \tau_{AB} \)) or the effect of changing \( \mu \) (which affects both \( \theta \) and \( \mu \tau_{AB} \)). From equation (2), it is clear that the probability that two sequences are identical decreases as \( \tau_{AB} \), \( N_\mu \), or \( \mu \) (and therefore \( \theta \)) increases, where the effect of \( \mu \) is stronger than the effect of \( N \). For example, changing \( \mu \) by one order of magnitude influences the probability of identical sequences more than changing \( N \) by one order of magnitude. These results can be observed in Figure 1, which depicts the probability that there exists at least one alignment in which a pair of species has zero divergence as a function of the number of loci and either \( \tau_{AB} \), \( N_\mu \), \( \mu \), or alignment length. From Figure 1, \( N \) has remarkably little influence over typical values (\( N \geq 10^3 \)).

**RESULTS**

**Simulations: Estimated Coalescence Times and Bias**

Figure 2 displays the results from simulating 11,000 gene trees at five different alignment lengths from the species tree \( (A,B),(C,D) \) where all branches had length \( x \), with either \( x = 5N \) or \( x = 0.5N \) generations, for different combinations of \( N \) and \( \theta \). The first column of Figure 2 shows the distributions of the maximum error for estimated coalescence times, \( D_{\text{max}}^{y} \). For a gene tree to be estimated accurately enough to satisfy condition (1), \( D_{\text{max}}^{y} \) must be less than \( m/2 \), where \( m \) is the shortest branch of the species tree. Thus we have \( m/2 = 2.5 \) and \( m/2 = 0.25 \) depending on whether the species tree branch lengths are 5 or 0.5 coalescent units, respectively. Points in the distributions above the horizontal dotted reference lines therefore correspond to gene trees estimated too inaccurately to satisfy condition (1). Under optimistic conditions, for example, tree height of \( 10N \) and \( \theta = 0.01 \),
the proportions of gene trees failing to be accurate enough at 500 nt and 1000 nt were 12.9% and 2.0%, respectively. For other combinations of tree height and θ, most gene trees fail to be estimated accurately enough for condition (1) to be satisfied for alignments of length 500 or 1000 nt. At a tree height of 1N, θ=0.001, and alignments of length 500 nt, none of the 11,000 simulated gene trees were estimated accurately enough.
to measure the bias, that is, the average value of 2

A

100-locus data sets (Fig. 2, third column). The true simulated gene trees per alignment length, using 100 trees, we applied STEM1.1 and STEM2.0 to the 11,000 time estimates in STEM from using estimated gene underestimates species divergence times.

Figure 2 exhibit greater variation than true coalescence error. Typically, the average estimated coalescence times are quite close to the average true coalescence time of 2x + 1N; however, estimated coalescence times in Figure 2 exhibit greater variation than true coalescence times, with distributions of estimated coalescence times approaching the distributions of true coalescence times as the alignment length increases. What is particularly striking is that estimated coalescence times can be drastically lower than true coalescence times and even speciation times. Thus, despite the observation that coalescence times overestimate species divergence times on average (Edwards and Beerli 2000; Jewett and Rosenberg 2012), a problem with using minimum estimated coalescence times is that they are likely to underestimate species divergence times.

To examine the effect on the bias in species divergence-time estimates in STEM from using estimated gene trees, we applied STEM1.1 and STEM2.0 to the 11,000 simulated gene trees per alignment length, using 100 replicates of 10-locus data sets and 100 replicates of 100-locus data sets (Fig. 2, third column). The true divergence time of 2x generations was subtracted from the estimated divergence between species A and C to measure the bias, that is, the average value of bias = \( \overline{D}_{AC} - 2x \) is reported. Because the expected minimum of 100 independent, identically distributed (i.i.d.) random variables is necessarily lower than the expected minimum of 10 i.i.d. random variables, curves for estimated species divergences are lower for 100 loci than for 10 loci. This result is not a problem for known gene trees as known coalescence times always overestimate species divergences. The theoretical bias for known gene trees is \( 1/L \), where \( L \) is the number of loci, which is approached as the alignment length increases. Thus, on known gene trees, the bias decreases as more loci are used. Because estimated coalescence times often underestimate species divergence times, however, this leads to more bias in STEM trees based on 100 loci compared with 10 loci using estimated gene trees for most settings used in the simulation, even when STEM has no trouble estimating the species-tree topology. Note that the setting in which 2x = 10N and \( \theta = 0.01 \) is particularly unchallenging in the sense that 98.5% of the 55,000 estimated gene trees matched the species-tree topology, and 100% of STEM1.1 and STEM2.0 trees matched the species-tree topology.

Further trends from the plots show that STEM1.1 and STEM2.0 had different amounts of bias only for combinations of alignment length, tree height, and value of \( \theta \) for which coalescence times of zero were estimated. The fact that STEM2.0 does not utilize coalescence times of zero (as long as at least one locus has a nonzero coalescence time), often leads to less bias for STEM2.0 than for STEM1.1. A different pattern emerged in the setting in which 2x = 1N and \( \theta = 0.001 \) from the other cases. In this case, because of the low level of \( \theta \), a 500-nt alignment with one pairwise difference between a and c overestimates the species divergences, and a 500-nt alignment with no pairwise difference between a and c leads to an estimate of zero for the coalescence time, thus understimating the species divergence. Because it is likely for at least one alignment out of 10 to have such an underestimate, STEM1.1 usually reported zero for the estimated species divergence between A and C. However, it also likely for at least one alignment to have at least one pairwise difference. In this case, STEM2.0 ignores the underestimates and only utilizes overestimates of the species divergences, generating an upward bias. This behavior can also be understood by looking at the distribution of estimated coalescence times for 500 nt, in which the species divergence time of 1N was never estimated, but either over- or underestimated due to the discrete number of mutations (Fig. 2).

Finally, we note that although true coalescence times, and therefore minimum true coalescence times from multiple loci, overestimate species divergences, situations in which minimum estimated coalescence times overestimate species divergences may be limited to settings in which there are few loci or in which alignment lengths are long. It is thus not clear to what extent correcting the upward bias in known coalescence times—e.g., Jewett and Rosenberg (2012)—would be useful in practice. However, because mean estimated coalescence times are quite close to mean true coalescence times, even for short sequences (Fig. 2), methods that correct for the upward bias in mean coalescence times, such as iSTEC (Helmkamp et al. 2012), are likely to be useful when applied to both estimated and true gene trees.

**Simulations: Inferring Species Trees**

Results are plotted in Figures 3 and 4 which illustrate RF distances using ML gene trees on all methods as well as Bayesian gene trees for STEAC and STEM1.1. Additional results with Bayesian gene trees on other methods, as well as false positive and false negative rates for STEM1.1 and STEM2.0 are in Supplementary Figures S2–S7.
Overall, the best-performing methods were STEAC using Bayesian gene trees and topology-based methods MP-EST, STAR, and STELLS using ML gene trees, with STELLS performing less well than the other topology-based methods for the lower mutation rate of $\theta = 0.001$, and STELLS performing slightly better than other methods for $\theta = 0.01$ with three alleles sampled per species. STEAC had similar, but slightly worse performance than the topology-based methods for $\theta = 0.01$. For $\theta = 0.001$, STEAC had dramatically different performance depending on whether ML or Bayesian gene trees were used, with better accuracy when applied to Bayesian gene trees. For $\theta = 0.001$ with ML gene trees, STEAC did worse than both versions of STEM for small numbers of loci, although STEAC did somewhat better than STEM with more than 10 loci using Bayesian gene trees (Fig. 4). The accuracy of STEAC increased steadily with the number of loci for all simulation settings. Below we discuss the effects of particular variables in the simulation settings, with particular emphasis on STEM.

**STEM1.1 versus STEM2.0.**—There were striking differences between STEM1.1 and STEM2.0 given the same ML input gene trees. In comparison with
STEM1.1 and STEM2.0 returned more resolved species trees for both values of $\theta$ (Supplementary Figs. S2 and S3) and for both one and three alleles sampled per species (Supplementary Fig. S4). For alignment lengths of 1000 nt, results from STEM1.1 and STEM2.0 were much more similar when $\theta = 0.01$ (Supplementary Fig. S2), particularly with 100 loci or fewer. These results are consistent with the observation that alignments of length 1000 nt are much less likely to result in unresolved gene trees than shorter alignments. For $\theta = 0.001$, unresolved gene trees are still likely with 1000 nt alignments, leading to different behaviors for STEM1.1 and STEM2.0 (Supplementary Fig. S3).

For most simulations, STEM2.0 had similar or lower RF distances than STEM1.1, particularly with larger numbers of loci and for three alleles sampled per species at any number of loci. Although the RF distances tended to be lower, STEM2.0 sometimes had higher false negative rates at higher numbers of loci. In every case considered, the false positive rates for STEM1.1 decreased or stayed the same with increasing numbers of loci; whereas for STEM2.0, the false positive rate sometimes increased for $\theta = 0.001$.

**Effect of ML versus Bayesian gene trees.**—STEM1.1 was more accurate using Bayesian consensus gene trees than for ML gene trees for $\theta = 0.01$, but returned highly unresolved trees for $\theta = 0.001$ for both types of gene trees. The results for $\theta = 0.01$ agree with Knowles et al. (2012) who also found either no difference or slightly better performance for STEM with Bayesian consensus gene trees. STEM was also more accurate using Bayesian gene trees when $\theta = 0.001$, whereas for $\theta = 0.01$, accuracy was approximately the same for both types of gene trees. However, for MP-EST, STAR, and STELLS, accuracy was higher using ML trees, particularly for $\theta = 0.001$ and for smaller numbers of loci. We note that although differences in performance using ML versus Bayesian gene trees tended to increase with an increasing number of loci for STEM1.1, these differences tended to decrease for methods other than STEM1.1.

STEM1.1 tended to infer more resolved species trees using Bayesian gene trees than using ML gene trees, perhaps due to the prior used in Bayesian estimates of gene trees generating positive lengths separating taxa even when they have identical sequences. A mutation rate of $\theta = 0.01$ resulted in STEM1.1 applied to Bayesian gene trees having similar behavior as STEM2.0 applied to ML gene trees. Although STEM1.1 had lower RF values using Bayesian rather than ML gene trees, it sometimes had higher false positive rates for 50 or more loci with $\theta = 0.01$. For $\theta = 0.001$, however, the behavior of STEM1.1 applied to Bayesian gene trees was more similar to STEM1.1 than to STEM2.0 applied to ML gene trees.
Effect of alignment length.—Not surprisingly, all methods tended to perform better on all measures with longer alignments (i.e., 1000 nt). For methods other than STEM1.1, performance was similar using empirical (with a mean of 550 nt) and constant 550 nt alignment lengths. For both one and three alleles sampled per species, STEM1.1 using ML gene trees often performed better with constant alignment lengths than with the empirical distribution of alignment lengths in terms of RF distance and false negative rate for $\theta = 0.01$, whereas the false positive rate exhibited little change (Figs. 3 and Supplementary S2). Methods other than STEM tended to be less affected by alignment length for $\theta = 0.01$; however, STELLS and STEAC were strongly affected by alignment length for $\theta = 0.001$. In particular, for $\theta = 0.01$ and one allele sampled per species, methods other than STEM did better with doubling the number of 550 nt alignments versus using half as many 1000 nt alignments. However, for $\theta = 0.001$ on ML gene trees, this pattern was reversed for STEAC and STAR. For example, for $\theta = 0.001$, using 50 1000-nt alignments resulted in substantially lower RF distances than 100 550-nt alignments, even though less sequence data was used (Fig. 3).

Although the RF distance for STEM2.0 appeared to be less affected by whether the alignment length was constant for $\theta = 0.01$, species trees were less resolved using constant alignment lengths, resulting in different false positive and false negative rates for the two alignment-length settings even when the RF distance was approximately the same (Supplementary Fig. S2). For $\theta = 0.001$ (Supplementary Fig. S3), STEM2.0 also returned fewer resolved species trees under the setting with constant alignment lengths. This lower resolution for constant alignment lengths is potentially due to more ties in the branch-length distribution for the inferred gene trees (as the same number of pairwise differences between two sequences would most likely occur for different alignment lengths in the empirical distribution). Methods other than STEM were less affected by whether alignments were variable or constant in length.

Effect of $\theta$.—The value of $\theta$ clearly had a major impact on all species tree methods, with greater accuracy for $\theta = 0.01$ relative to $\theta = 0.001$. STEM1.1 mostly returned completely unresolved species trees when $\theta = 0.001$, yet mostly returned fully resolved species trees at smaller numbers of loci for $\theta = 0.01$. For $\theta = 0.001$, STELM2.0 had similar or worse RF distances as STEM1.1 due to trees that were more resolved but with high false positive rates (Supplementary Fig. S2). We note that by fixing the species tree height in terms of coalescent units and changing $\theta$, we are essentially modifying only the mutation rate, $\mu$. As a consequence, the probability of identical sequences in an alignment is greatly increased with $\theta = 0.001$ compared with $\theta = 0.01$. Methods other than STEM also performed much worse with $\theta = 0.001$, but still had steadily decreasing RF distances as the number of loci increased. Of the non-STEM methods, STEAC using ML gene trees and STELLS were the most sensitive to the small value of $\theta$.

Effect of the number of alleles.—All methods other than STEM1.1 did better with three alleles for a fixed number of loci. For STEM1.1 applied to ML gene trees, using three alleles sampled per species resulted in lower RF distances than using one allele sampled per species for small numbers of loci (Supplementary Figs. S2 and S4). However, when the alignment length was less than 1000 nt and there were three sampled alleles per species, the false positive rate was lower and the false negative rate was higher, and the inferred species tree became less resolved more quickly with increasing numbers of loci at all alignment lengths. RF distances were also larger for three alleles sampled per species when there were more than roughly 30 loci and alignments were less than 1000 nt. For 1000-nt alignments, RF distances were smaller for fewer numbers of loci using STEM1.1, but were increasing with the number of loci when there were more than about 50 loci using three alleles sampled per species, whereas RF distances were decreasing with the number of loci using one allele sampled per species. Consequently, RF distances were similar with 500 loci. Generally, false positive rates decreased to zero more quickly using three alleles sampled per species than one allele sampled per species, which is consistent with the inferred species trees being less resolved when three alleles were sampled per species. In contrast, STEM2.0 always performed better with three alleles than with one allele sampled per species by any measure.

Primate Analysis

Gene trees estimated with ML were fully resolved for 6376 out of 8881 (71.8%) alignments. Of these, 3969 (44.7%) of the 8881 alignments resulted in fully resolved gene trees that matched the assumed species tree with chimpanzees and humans forming a clade. Results were similar for Bayesian gene trees, with 6499 out of 8881 (73.1%) of gene trees being fully resolved and 4101 (46.2%) fully resolved gene trees matching the species tree. Although matching gene trees were inferred more often using MrBayes, false positives were also slightly more common for Bayesian gene trees. For example, the (C.C) clade was inferred 1128 times using ML and 1186 times with MrBayes, and similarly for other incorrect clades. Alignment lengths for the data averaged ~501 nt (standard deviation (s) ~159) for the 8881-locus data set, ~550 nt (s ~113) for the 7582-locus data set with alignment lengths greater than or equal to 300, and ~652 nt (s ~40) for the 3171-locus data set with alignment lengths greater than or equal to 600 nt. The proportion of variable sites over each data set was ~8.9%, which is a typical level of variation for multilocus phylogenetic studies.

To visualize the amount of variation in the data, the number of segregating sites, $S$, and numbers of
pairwise distances \( \pi_d \) for macaque-orangutan, human-chimp, and chimp-gorilla are plotted in Supplementary Figure S8. Although these statistics increase linearly (on average) with the alignment length, there is quite a bit of variation, and we observe that undiverged sequences occur even with sequences over 600 nt in length. The longest alignment in the 8881-locus data set for which chimpanzee and gorilla have identical sequences is 665 nt long (Supplementary Fig. S8a), which when included forces STEM1.1 with ML gene trees to place the divergence time of chimpanzee and gorilla as zero in the species tree.

Results for the estimated species trees under various conditions (minimum alignment length, ML versus Bayesian gene trees) are provided in Table 1 for STEM1.1 and STEM2.0. As expected, increasing the minimum alignment length increases the resolution of the estimated species tree. STEM2.0 applied to ML gene trees and STEM1.1 applied to Bayesian gene trees also result in more-resolved species-tree estimates. For species trees that were resolved, however, orangutan was often incorrectly inferred to be the outgroup (for alignments at least 600 nt in length) or incorrectly grouped with macaque (for STEM2.0 analyses that included alignments less than 600 nt). These incorrect placements of the outgroup are problems with rooting the species tree, and consequently no misleading groups on unrooted versions of the species trees were inferred for any of the analyses. Results for smaller sets of loci are shown in Figure 5.

Table 1. Estimated species tree topologies for the primate data set using different alignment lengths and methods

<table>
<thead>
<tr>
<th>Alignments</th>
<th>Method</th>
<th>Species tree estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>8881 (all)</td>
<td>STEM1-ML</td>
<td>(C, G, H, M, O)</td>
</tr>
<tr>
<td>8881 (all)</td>
<td>STEM2-ML</td>
<td>((C, H), (G, M, O))</td>
</tr>
<tr>
<td>8881 (all)</td>
<td>STEM1-Bayes</td>
<td>((C, H), G, M, O)</td>
</tr>
<tr>
<td>7582 (≥300 nt)</td>
<td>STEM1-ML</td>
<td>(C, G, H, M, O)</td>
</tr>
<tr>
<td>7582 (≥300 nt)</td>
<td>STEM2-ML</td>
<td>((C, H), (G, M, O))</td>
</tr>
<tr>
<td>3171 (≥600 nt)</td>
<td>STEM1-ML</td>
<td>((C, G, H), M, O)</td>
</tr>
<tr>
<td>3171 (≥600 nt)</td>
<td>STEM2-ML</td>
<td>((C, H), (G, M, O))</td>
</tr>
<tr>
<td>3171 (≥600 nt)</td>
<td>STEM1-Bayes</td>
<td>((C, H), G, M, O)</td>
</tr>
</tbody>
</table>

FIGURE 5. Robinson-Foulds (RF) distances returned by STEM1.1, STEM2, STELLS, MP-EST, STEAC, and STAR using maximum likelihood (ML) and STEM1.1 using Bayesian gene trees on random subsets of loci from the 8881-locus ape data set. The first column is the full data set. The second column is alignments of length at least 300 nt. The third column is alignments of length at least 600 nt.

STEM1.1 versus STEM2.0.—Overall, STEM1.1 tended to produce less-resolved species trees than STEM2.0, particularly for larger numbers of loci, with the exception of alignments with length at least 600 nt, for which all methods produced highly resolved species trees at any number of loci.

On the original 8881-locus data set, as the number of loci increases, the species tree estimates become increasingly unresolved for STEM1.1 applied to ML gene trees (Supplementary Fig. S9). This pattern is expected if there exists a reasonable chance that a pair of sequences is undiverged at a locus, which occurs for this data set. Consequently, though the RF distance is increasing as the number of loci increases, the number of false positives is decreasing and the number of false negatives is increasing. At lower numbers of loci (e.g., 10 and 20), the false positive and false negative rates contribute roughly equally to the RF distance. However, for larger numbers of loci, almost all of the RF distance is due to the false negative rate, yielding species-tree estimates that are uninformative but not misleading. The behavior of STEM2.0 on the 8881-locus data set is quite different from

MP-EST, STAR, STEAC, and STELLS.—Methods other than STEM correctly estimated the fully resolved species tree between 88% and 91% of the time on 10 loci on the full 8881-locus data set for any combination of Bayesian versus ML gene trees, alignment length, and method. The accuracy improved to 97–99% for 20 loci, and 100% for more than 60 loci, with accuracy. Differences between methods were therefore minor for this data set and all had much lower RF values than STEM.
STEM1.1, with STEM2.0 having roughly the same (fairly high) probability of returning a fully resolved species tree at any number of loci. There exists a slight upward trend in the RF distance, false positive rate, and false negative rate at larger numbers of loci, which is largely due to an influential locus, alignment 2111 (592 nt with six polymorphic sites) in the 14,663-locus data set, which has only one pairwise difference between orangutan and macaque and causes the STEM2.0 tree to place these two taxa together (Table 1). Removing alignment 2111 from the 8881-locus data set results in STEM2.0 returning the correct species-tree topology; however, locus 2111 does not affect STEM1.1 because there are other loci that force the estimated divergence to be even lower (i.e., zero) between these taxa, but which STEM2.0 does not use. Similarly, alignment 2111 does not affect STEM1.1 applied to Bayesian gene trees, as there exist other loci with smaller (although non-zero) divergences that are used in the species-tree estimate.

Effect of ML versus Bayesian gene trees. — For MP-EST, STAR, STEAC, and STELLS, differences between ML and Bayesian gene trees were fairly minor. When gene trees were estimated with MrBayes, STEM1.1 returned more resolved species trees than with ML gene trees, with results being similar to STEM2.0 applied to ML gene trees for up to 100 loci. For larger numbers of loci, STEM1.1 applied to Bayesian gene trees tended to be less resolved than STEM2.0 applied to ML gene trees, generating results for STEM1.1 applied to Bayesian gene trees intermediate between STEM1.1 and STEM2.0 applied to ML gene trees. STEAC was also slightly more accurate on Bayesian gene trees, although the difference was not statistically significant. For topology-based methods, species trees estimated from Bayesian consensus gene trees tended to have slightly higher RF values than from ML gene trees for relatively small numbers of loci (results not shown).

Effect of alignment length. — The 8881-locus alignment data set has many short loci, with 14.6% of loci having fewer than 300 nt, and 64.3% of loci having fewer than 600 nt. To investigate the effect of alignment length, we considered subsetting the data to have only alignments with at least 300 nt and 600 nt each (Supplementary Fig. S9). For alignments at least 300 nt in length, STEM1.1 applied to ML gene trees had more resolved species tree estimates than for the full data set. STEM1.1 applied to ML gene trees also had smaller RF distances and lower false negative rates for small numbers of loci; however, it had higher false positive rates for large numbers of loci. Unfortunately, there exists another influential locus, alignment 5403, which has only one polymorphic site, a singleton with the mutation on the orangutan sequence, which tends to place orangutan as an outgroup. However, this alignment has no influence on STEM1.1 applied to Bayesian gene trees or on STEM2.0 applied to ML gene trees. Finally, we considered alignments with at least 600 nt (Supplementary Fig. S9, third column), which excludes alignment 2111 (592 nt) and alignment 5403 (368 nt). For STEM1.1 applied to ML gene trees, the results are quite similar to the results for alignments of at least 300 nt in length due to other loci that place orangutan as an outgroup. Furthermore, orangutan was often placed as an outgroup for STEM2.0 and STEM1.1 applied to Bayesian gene trees as well, leading to somewhat worse performance for these longer alignments than for the data sets that included shorter alignments.

**DISCUSSION**

**Maximum Likelihood Versus Bayesian Gene Trees**

An interesting pattern was that the topology-based methods—MP-EST, STAR, and STELLS—performed better using ML gene trees, whereas STEM1.1 and STEAC, both based on coalescence times, tended to have the same or better performance with Bayesian rather than ML gene trees (see also Knowles et al. 2012). This pattern was observed for both the simulations and the empirical primate data. The effect is less clear for STEM2.0 (which is equivalent to STEM1.1 on Bayesian gene trees), because STEM2.0 did slightly better with ML trees than STEM1.1 with Bayesian trees for the empirical data but slightly worse on simulated data. As inferring species trees from gene trees become more common, it is worthwhile to investigate the impact of methods for inferring the gene trees on the species tree estimates. The results from both the empirical data analysis and the simulations display an interesting similarity between STEM1.1 applied to Bayesian gene trees and STEM2.0 applied to ML gene trees. For finite sequences, Bayesian gene trees do not contain estimated coalescence times of zero, even for pairs of sequences that are undiverged, due to the influence of the prior on branch lengths. Interpreting coalescence times as distances between sequences (as is done for GLASS), it may appear odd that two identical sequences should have a distance other than zero. If there is prior information that the sequences are sampled from distinct species, however, then it might be reasonable to provide undiverged sequences a nonzero distance. On the other hand, if species assignment is uncertain, then a prior that always assigns nonzero distances to undiverged sequences might be undesirable. Similarly, the strategy of STEM2.0 of ignoring zero-divergences between different taxa is reasonable if the taxa are known to be from distinct species. However, this procedure might not be reasonable if species assignment is unknown, particularly when STEM is used for species delimitation as in the program SpeDeSTEM (Ence and Carstens 2011), which currently uses STEM1.1. STEM2.0 applied to ML gene trees can be thought of as employing an alternate strategy for dealing with undiverged sequences, namely essentially ignoring them. Although the justification for ignoring undiverged
sequences when estimating species trees using ML is unclear, it is reasonable to believe that coalescence times of zero between lineages sampled from distinct species are underestimates and must be incorrect. Thus, it might be reasonable to ignore these observations, though there exists some risk in increasing the bias for the estimated species divergence for small $\theta$, short alignments, and recent species divergences (e.g., Fig. 2).

Although STEM2.0 applied to ML gene trees and STEM1.1 applied to Bayesian gene trees are two different strategies for dealing with underestimated coalescence times, the two approaches often lead to fairly similar RF distances between estimated and true species trees.

Comparison with Previous Studies

Previous simulation studies using STEM1.1 and GLASS have illustrated a dramatic difference when these methods are applied to known gene trees versus estimated gene trees (Liu et al. 2009; Huang et al. 2010; Wu 2012). In Liu et al. (2009), the accuracy of GLASS when the molecular clock is satisfied does not appear to change much with an increasing number of loci. In the study by Huang et al. (2010), the RF distance always improved with an increasing number of loci using at most 50 loci. In McCormack et al. (2009), the accuracy at 1N tree depth improved up to 27 loci, but then decreased from 57 to 100 loci, although the change in direction was statistically similar. Significant, our results, under identical sampling designs in McCormack et al. (2009) and Huang et al. (2010), this nonmonotonic relationship between accuracy and number of loci was observed when gene trees were estimated with ML (McCormack et al. 2009) but not with MrBayes (Huang et al. 2010). With multiple alleles sampled per species, both McCormack et al. (2009) and Huang et al. (2010) found increased accuracy with three alleles compared with one allele sampled per species for fixed numbers of loci; however, this sampling scenario was tested with at most nine loci. We also observed increased accuracy with three alleles at 10 loci on STEM1.1 for $\theta = 0.01$, however, by 30 loci, sampling only one allele was statistically significantly more accurate than sampling three alleles in terms of RF distance (although not in terms of the number of false positives).

Two studies using more loci have found that accuracy of STEM applied to estimated ML gene trees can decrease with the number of loci. The study by Hird et al. (2010) also found a nonmonotonic relationship between accuracy of STEM1.1 applied to ML gene trees and the number of loci, using both RF and branch-score distance. The decrease in accuracy was more pronounced in Hird et al. (2010) with one individual sampled per species, but to a lesser extent also appeared with three individuals per species. The study by Hird et al. (2010), similar to our article, also considered estimating the great ape species tree, but included two populations each of orangutan and bonobos, and used multiple sampled alleles per species. In their study, a 10-locus data set was used, and they found that the correct species tree was observed regardless of which subsets of alleles per species were used. The Hird et al. (2010) analysis was based on data in Fischer et al. (2006), which had alignments ranging from 650 to 1500 nt in length. For our great ape analyses using subsets of 10 randomly selected loci, the correct, fully resolved species tree was estimated in 575 out of 1000 replicates using all alignments, and in 730 out of 1000 replicates using alignments at least 600 nt in length. The results in Hird et al. (2010) based on a single set of 10 loci might not hold for some other sets of alignments from these taxa, particularly for the alignment lengths found in the Burgess and Yang (2008) metanal data set.

Interestingly, Leaché and Rannala (2011) found that varying the mutation rate $\theta$ had little effect on the accuracy of species trees estimated by STEM. In contrast, we observed that there was a large impact on changing $\theta$ when the height of the species tree was kept constant in terms of coalescent units. Part of the difficulty in interpreting the effect of $\theta$ is that it depends on two parameters, $N$ and $\mu$. Our theoretical results illustrate that these two parameters have different effects on the probability that pairs of sequences are undiverged. Thus, the impact of changing $\theta$ on STEM1.1 may partly depend on whether $N$ or $\mu$ is changed. The effect of $\theta$ in Leaché and Rannala (2011) is observed by fixing the species tree at height $T$, where $T$ is measured as generations multiplied by mutation rate $\mu$. The height of the species tree in coalescent units is therefore proportional to $T/\theta$. Thus, keeping $T$ fixed and varying $\theta$ (or vice versa), changes the height of the species tree in coalescent units. Because $T$ is fixed, varying $\theta$ can be achieved by keeping $\mu$ fixed but varying the population size $N$, which we expect to have less of an influence than varying the mutation rate based on our observations from equation (2) and Figure 1. In contrast, if the species tree height measured in generations (or coalescent units) is fixed, and $\mu$ varies, then both $\theta$ and $T$ change. From the results in Leaché and Rannala (2011), if $\mu$ is increased whereas $T/\theta$ is fixed (thus increasing both $T$ and $\theta$), then the performance of STEM1.1 improves more dramatically than the other methods tested in Leaché and Rannala (2011).

Our results are consistent with the previous studies (McCormack et al. 2009; Hird et al. 2010; Leaché and Rannala 2011; Wu 2012) in observing that the accuracy of STEM applied to estimated ML gene trees can decrease with the number of loci. The results from the ape data analysis confirm in an empirical data set that increasing the number of loci can increase the probability that the GLASS tree will not match the species tree, although
in some cases, particularly with STEM1.1 applied to ML gene trees, this estimation error may be due to the inferred species tree being increasingly likely to be at least partially unresolved. This observation highlights the importance of considering false positive and false negative rates, as well as the RF distance (Swenson et al. 2011). The great ape data also shows that genes that are highly influential for STEM are not necessarily influential for other methods, a finding that is difficult to observe from the simulations.

Our hope in this article has been to help understand the relatively poor performance of STEM in comparison to methods such as MP-EST, STAR, STEAC, and STELLS despite the desirable properties of GLASS applied to known gene trees and apparent theoretical justification for applying GLASS on estimated gene trees (Mossel and Roch 2010). In particular, we have demonstrated that much of the problem appears to be due to undiverged sequences or ties in estimated divergence times, and that the decrease in accuracy is often due more to increasing false negative rates rather than increasing false positive rates, particularly when using ML gene trees.

Although topology-based methods require large sample sizes to infer difficult species trees (Degnan 2013), sufficiently large data sets are becoming available increasingly often (e.g., Song et al. 2012; McCormack et al. 2013a,b). Among methods that can handle such phylogenomic data sets, we have shown that topology-based methods offer robustness to undiverged sequences and can have similar or better performance on phylogenomic data sets as some current methods using coalescence times.

**APPENDIX A**

Proof of Proposition 1. Let the probability of no mutation at a site within time $t>0$ be $p(t)$, with $0<p(t)<1$. The probability that two particular sequences are identical at a site is at least $|p(t)|^2 > 0$, where $t$ is the coalescence time between the lineages, since two sequences might also be identical due to back mutations or convergent mutations. (Here $|p(t)|^2$ is the probability that neither sequence had a mutation at the site.) Under assumptions 1–6, the probability $p_0$ that a pair of sequences $i$ (from species $A$) and $b$ (from species $B$) of length $y \leq Y$ are identical is

$$p_0 = \int_0^\infty \left( |p(t)|^2 \right)^y \frac{p(t)dt}{t} \geq 0.$$ 

Because $p_0 < 1$, the probability that a sample of $z$ loci has at least one locus $i$, with $\tilde{D}_{ab}^0 = 0$ is $1-(1-p_0)^z \rightarrow 1$ as $z \rightarrow \infty$. Thus,

$$|\tilde{D}_{AB} - D_{AB}| \rightarrow D_{AB} \tau_{AB} \geq m \rightarrow m \frac{m}{2}.$$ 

Therefore condition (1) fails for any pair of species $A$ and $B$. Repeating this argument for all pairs of species, the estimated divergence time $\tilde{D}_{AB} = \min \tilde{D}_{ab}^0$ between all pairs of species $A$ and $B$ goes to zero as $z$ goes to infinity.

**APPENDIX B**

Under the $N_r$ model, the probability of a mutation within length $t$, where $t$ is time measured in generations, is (Tuffley and Steel 1997)

$$p(t) = \frac{r-1}{r} \left( 1 - e^{-\mu t} \right),$$ 

where $\mu$ is the per-site per-generation mutation rate. The probability of the site pattern $\times \times$ under the $N_r$ model as a function of time, can be obtained by considering the unrooted two-taxon gene tree with a branch of length $2t$ generations. Thus the probability that the site pattern is $\times \times$ is

$$\Pr[\tilde{D}_{ab}^0 = 0|t] = 1 - p(2t) = \frac{1}{r} \left( 1 - \frac{1}{r} \right) e^{-\frac{2\mu t}{r}}.$$ (A1)

The probability that a pair of sequences of length $y$ are undiverged, given that they evolved on the same gene tree of length $t$, is

$$\Pr[\tilde{D}_{ab}^0 = 0|t] = \frac{1}{t} + \left( \frac{1}{r} - \frac{1}{r} \right) e^{-\frac{2\mu t}{r}} \left[ \sum_{i=0}^{y-1} \left( \frac{y}{i} \right) (1 - \frac{1}{r}) e^{-\frac{2\mu t}{r}} \right]^{-y-i}. $$ (A2)

which is obtained by applying the binomial formula.
The probability density of the coalescence time for a pair of lineages randomly sampled from two distinct species is

$$f(t) = \frac{1}{N^2} \frac{e^{-t/\tau_{AB}}}{\tau_{AB}} 1_{[0, \tau_{AB}]}$$  \hspace{1cm} (A3)$$

where $t$ and $\tau_{AB}$ are measured in generations. Integrating the probability that two sequences are undiverged (equation (A2)) with respect to the probability density in equation (A3) over all possible coalescence times (from $\tau_{AB}$ to $\infty$) yields equation (2).

REFERENCES

Blair J. E., Coffey M. D., Martin F. N. 2012. Species tree estimation for the late blight pathogen, Phytophthora infestans, and close relatives. PLoS ONE 7:e37033.


