JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 6, 2012
© Mary Ann Liebert, Inc.
Pp. 632–649
DOI: 10.1089/cmb.2012.0042

Improvements to a Class of Distance Matrix Methods for Inferring Species Trees from Gene Trees

LAURA J. HELMKAMP,[1] ETHAN M. JEWETT,[2] and NOAH A. ROSENBERG[2]

ABSTRACT

Among the methods currently available for inferring species trees from gene trees, the GLASS method of Mossel and Roch (2010), the Shallowest Divergence (SD) method of Maddison and Knowles (2006), the STEAC method of Liu et al. (2009), and a related method that we call Minimum Average Coalescence (MAC) are computationally efficient and provide branch length estimates. Further, GLASS and STEAC have been shown to be consistent estimators of tree topology under a multispecies coalescent model. However, divergence time estimates obtained with these methods are all systematically biased under the model because the pairwise interspecific gene divergence times on which they rely must be more ancient than the species divergence time. Jewett and Rosenberg (2012) derived an expression for the bias of GLASS and used it to propose an improved method that they termed iGLASS. Here, we derive the biases of SD, STEAC, and MAC, and we propose improved analogues of these methods that we call iSD, iSTEAC, and iMAC. We conduct simulations to compare the performance of these methods with their original counterparts and with GLASS and iGLASS, finding that each of them decreases the bias and mean squared error of pairwise divergence time estimates. The new methods can therefore contribute to improvements in the estimation of species trees from information on gene trees.

Key words: algorithms, coalescence, phylogenetic trees.

1. INTRODUCTION

MANY METHODS EXIST FOR INFERRING SPECIES TREES FROM GENE TREES (Maddison, 1997; Ewing et al., 2008; Degnan and Rosenberg, 2009; Kubatko et al., 2009; Liu et al., 2009; Than and Nakhleh, 2009). Among these methods, several are now available that are computationally efficient and capable of estimating branch lengths. This collection of methods includes the GLASS method of Mossel and Roch (2010), which was developed independently by Liu et al. (2010) under the name Maximum Tree, the Shallowest Divergence (SD) method of Maddison and Knowles (2006), and the Species Tree Estimation using Average Coalescence times (STEAC) method of Liu et al. (2009).

GLASS, SD, and STEAC are all distance-matrix methods. Such methods first estimate an evolutionary distance between each pair of taxa and then construct a species tree that provides an approximate

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan.
[2]Department of Biology, Stanford University, Stanford, California.

representation of the pairwise distances. In GLASS, SD, and STEAC, divergence times are estimated for each pair of species, and a hierarchical clustering method is then applied to the pairwise distances to construct an estimate of the species tree. Thus, each distance-matrix method can be viewed as an estimator of pairwise species divergence times combined with a hierarchical clustering procedure.

To understand the general approach by which GLASS, SD, and STEAC estimate the divergence time between a pair of taxa A and B, consider a set of L loci indexed by ℓ, and denote by $n_A^{(\ell)}$ and $n_B^{(\ell)}$ the numbers of lineages sampled at locus ℓ in taxa A and B, respectively. Each method takes as input a set of estimated coalescence times for all pairs of taxa. Denote the sets of lineages sampled from taxa A and B at locus ℓ by $\{a_i^{(\ell)}\}_{i=1}^{n_A^{(\ell)}}$ and $\{b_j^{(\ell)}\}_{j=1}^{n_B^{(\ell)}}$, and denote the true coalescence time between lineages $a_i^{(\ell)}$ and $b_j^{(\ell)}$ by $T_{a_i^{(\ell)}, b_j^{(\ell)}}$. We assume that an estimate $\hat{T}_{a_i^{(\ell)}, b_j^{(\ell)}}$ of $T_{a_i^{(\ell)}, b_j^{(\ell)}}$ has been obtained by some procedure that is left unspecified; the focus of GLASS, SD, and STEAC is not on the estimation of the $T_{a_i^{(\ell)}, b_j^{(\ell)}}$ themselves, but rather, on the way in which a collection of values of $\hat{T}_{a_i^{(\ell)}, b_j^{(\ell)}}$ is used to estimate species divergence times. Throughout this paper, we use the term *pairwise interspecific coalescence* to refer to the $T_{a_i^{(\ell)}, b_j^{(\ell)}}$; that is, the $n_A^{(\ell)} \times n_B^{(\ell)}$ coalescences between one of the $n_A^{(\ell)}$ lineages from taxon A and one of the $n_B^{(\ell)}$ lineages from taxon B at locus ℓ. We differentiate this concept from that of an *interspecific coalescent event*, at which multiple pairwise interspecific coalescences, as defined here, can simultaneously occur. For example, in the species tree depicted in each box of Figure 1, two interspecific coalescent events occur for the locus shown in blue. At each of these events, two pairwise interspecific coalescences occur. The gene tree at the other locus, in orange, contains three interspecific coalescent events. One pairwise interspecific coalescence occurs at each of the first two interspecific coalescent events, and two pairwise interspecific coalescences, between the rightmost lineage from the right taxon and each lineage from the left taxon, occur at the third interspecific coalescent event. From these $n_A^{(\ell)} \times n_B^{(\ell)}$ pairwise interspecific coalescences, we denote the estimated minimum pairwise interspecific coalescence time between the taxa at locus ℓ by $\check{T}_{AB}^{(\ell)} = min_{i,j}(\hat{T}_{a_i^{(\ell)}, b_j^{(\ell)}})$, and the estimated mean pairwise interspecific coalescence time by $\bar{T}_{AB}^{(\ell)} = \frac{1}{n_A^{(\ell)} n_B^{(\ell)}} \sum_{i=1}^{n_A^{(\ell)}} \sum_{j=1}^{n_B^{(\ell)}} \hat{T}_{a_i^{(\ell)}, b_j^{(\ell)}}$.

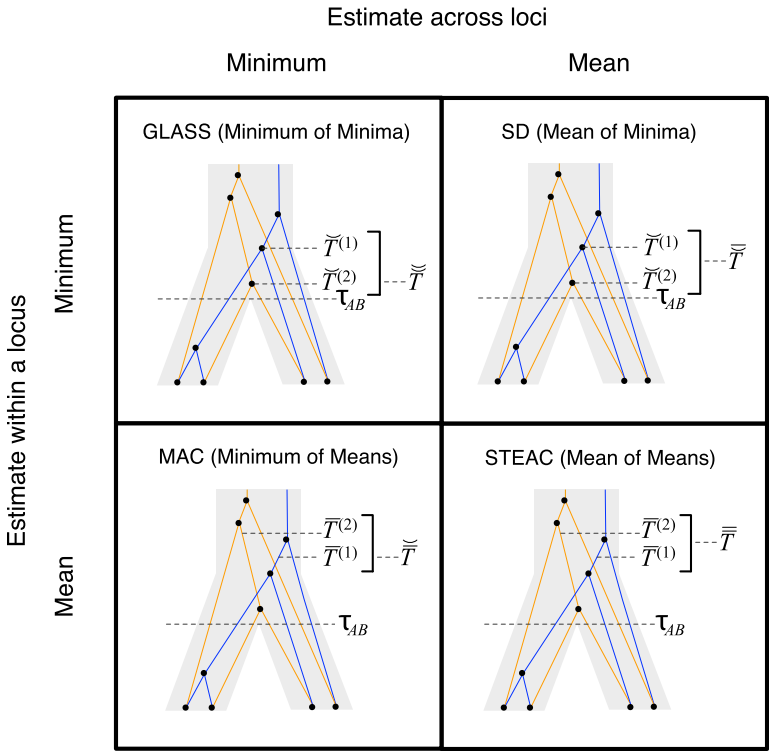


FIG. 1. Four methods for estimating divergence times from sets of multiple lineages sampled at many loci in two species. The same species tree and gene trees are pictured in all four panels. In the first row, we consider $\check{T}^{(1)}$, the minimum interspecific coalescence time at the first locus, in blue, and $\check{T}^{(2)}$, the minimum interspecific coalescence time at the second locus, in orange. Considering $\check{T}^{(1)}$ and $\check{T}^{(2)}$, we take either the minimum, resulting in the GLASS estimate $\check{\check{T}}$, or the mean, resulting in the SD estimate $\bar{\check{T}}$. In the second row, we consider the mean interspecific coalescence times at the first and second loci, $\bar{T}^{(1)}$ and $\bar{T}^{(2)}$, from which we take the minimum to obtain the MAC estimate $\check{\bar{T}}$ or the mean to obtain the STEAC estimate $\bar{\bar{T}}$. Note that in this example, locus 1 has a greater minimum interspecific coalescence time but a smaller mean interspecific coalescence time than locus 2.

The GLASS, SD, and STEAC estimation methods each consider a minimum or mean of either the $\check{T}_{AB}^{(\ell)}$ or the $\bar{T}_{AB}^{(\ell)}$ over all loci (Fig. 1). For example, the GLASS estimator of the divergence time T_{AB}, denoted $\check{\bar{T}}_{AB}$, is obtained by taking the minimum over loci of the locus-specific minimum pairwise interspecific coalescence times; that is, $\check{\bar{T}}_{AB} = min_{\ell}(\check{T}_{AB}^{(\ell)})$. The SD estimator of T_{AB} is obtained by taking the mean over loci of the locus-specific minimum pairwise interspecifc coalescence times, or $\bar{\check{T}}_{AB} = \frac{1}{L}\sum_{\ell=1}^{L} \check{T}_{AB}^{(\ell)}$. The STEAC estimator of T_{AB} is the mean over loci of the locus-specific mean pairwise interspecific coalescence times, or $\bar{\bar{T}}_{AB} = \frac{1}{L}\sum_{\ell=1}^{L} \bar{T}_{AB}^{(\ell)}$. We also introduce a fourth divergence time estimator equal to the minimum over loci of the locus-specific mean pairwise interspecific coalescence times, or $\check{\bar{T}}_{AB} = min_{\ell}(\bar{T}_{AB}^{(\ell)})$. We will call this method Minimum Average Coalescence (MAC).

The species divergence time estimates given by GLASS, SD, STEAC, and MAC are all systematically biased under the multispecies coalescent, a basic evolutionary model for describing the evolution of gene trees conditional on species trees (Degnan and Rosenberg, 2009). This bias arises because, under the model, true pairwise interspecific gene coalescence times $T_{a_i^{(\ell)}, b_j^{(\ell)}}$ always exceed the species divergence time. If $\hat{T}_{a_i^{(\ell)}, b_j^{(\ell)}}$ estimates $T_{a_i^{(\ell)}, b_j^{(\ell)}}$ perfectly, then because the estimates from GLASS, SD, STEAC, and MAC ($\check{\bar{T}}_{AB}, \bar{\check{T}}_{AB}, \bar{\bar{T}}_{AB}$, and $\check{\bar{T}}_{AB}$, respectively) all consist of means and minima of the $\hat{T}_{a_i^{(\ell)}, b_j^{(\ell)}}$, under the model, they too must exceed the true species divergence time.

For the GLASS method, Jewett and Rosenberg (2012) addressed the issue of bias by proposing an improved method for estimating pairwise divergence times. This method, which they called iGLASS, adjusts the GLASS estimate downward by an amount related to the bias of GLASS. Jewett and Rosenberg showed that iGLASS is consistent for estimating pairwise divergence times and that it can be combined with a suitable clustering algorithm to produce a consistent estimator of the species tree topology. Through simulations, they found that iGLASS greatly reduces the bias and mean squared error of pairwise divergence time estimates produced by GLASS.

Here, we propose improved analogues for SD, STEAC, and MAC, which we call iSD, iSTEAC, and iMAC, respectively, under which bias in the estimation of species divergence times is substantially reduced or eliminated. In Section 2, we derive the improved methods for estimating pairwise divergence times. Our derivations first quantify the bias in pairwise estimates of species divergence time for each of the original three methods as a function of the true divergence time, the number of lineages sampled from each taxon, and the number of sampled loci. Given a divergence time estimate obtained using one of the original methods, an improved estimate is produced by subtracting from the original estimate its bias; more precisely, because the bias for a given scenario depends on the true divergence time, which we are attempting to estimate, we calculate improved estimates by solving an equation for the true divergence time that will be detailed in Section 2. In Section 3, we evaluate the methods for estimating pairwise divergence times, and we compare the performance of the improved methods with each other and with their original counterparts. We also expand the analysis to trees with more than two taxa; in these cases, the improved version of a given method (GLASS, SD, STEAC, or MAC) is obtained by calculating the improved pairwise divergence time estimate for each pair of taxa and then applying a clustering method to the matrix of improved estimates. We compare the original and improved methods for inferring these larger trees with respect to both the bias and mean squared error in pairwise divergence time estimates and the proportion of sampled trees for which the correct topology is inferred.

2. DERIVATION OF IMPROVED METHODS

We follow the assumptions of a simple multispecies coalescent model (Degnan and Rosenberg, 2009). Each branch i of the species tree is assumed to have a constant effective population size N_i. Looking backward in time from the present at a species tree branch with n sampled lineages, we assume that all $\binom{n}{2}$ pairs of lineages are equally likely to coalesce and that the waiting time until coalescence occurs is exponentially distributed with mean $1/\binom{n}{2}$ coalescent time units (where one unit is defined as $2N_i$ generations). The species tree and gene trees are assumed to be binary. Furthermore, any two species or genetic lineages are assumed to be equidistant from their common ancestor in time units of years, so that species trees and gene trees are ultrametric. We also assume that $N_i = N$ for all i, and that generation times and mutation rates are also equivalent across species tree branches.

Bias reduction is carried out in a similar way for each of the four estimators that we consider. As we have discussed, Jewett and Rosenberg (2012) examined the bias reduction in the case of GLASS; here, we proceed more generally. Consider a specific estimator t_{AB} chosen from among the four in our study; that is, t_{AB} represents either $\breve{T}_{AB}, \bar{\breve{T}}_{AB}, \breve{\bar{T}}_{AB}$, or $\bar{\bar{T}}_{AB}$. To obtain an improved version of t_{AB}, let $\hat{t}_{AB}$ denote an observation of the estimator t_{AB}. We then obtain the improved estimate by finding the divergence time at which the expected value of the estimator t_{AB} under the model is equal to the observed estimate. That is, we solve

$$\hat{t}_{AB} = E_{\tau_{AB}}[t_{AB}] \tag{1}$$

for τ_{AB}.

In the case in which the estimate $\hat{t}_{AB}$ is smaller than the smallest possible value for $E_{\tau_{AB}}[t_{AB}]$, or $E_0[t_{AB}]$, it is not biologically meaningful to solve Equation (1), since the equation produces negative divergence time estimates. To circumvent this problem, define a function g as

$$g(\tau_{AB}) = E_{\tau_{AB}}[t_{AB}]. \tag{2}$$

The improved estimate for the pairwise species divergence time, $\hat{t}^*_{AB}$, is now defined piecewise:

$$\hat{t}^*_{AB} = \begin{cases} g^{-1}(\hat{t}_{AB}) & \text{if } \hat{t}_{AB} \geq E_0[t_{AB}] \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

As we will see, except in the case of STEAC, this procedure gives an estimate of τ_{AB} that is biased when more than one lineage is sampled in each taxon. However, the bias of each improved estimator is greatly reduced compared to that of the corresponding original estimator.

To simplify our derivations, we define a random variable V_{AB} as the difference between the estimator t_{AB} and the true speciation time τ_{AB}; that is, $V_{AB} = t_{AB} - \tau_{AB}$. V_{AB} measures the extent to which a random estimate exceeds the true speciation time. We can now express $E_{\tau_{AB}}[t_{AB}]$ from Equation (2) as $E_{\tau_{AB}}[t_{AB}] = \tau_{AB} + E_{\tau_{AB}}[V_{AB}]$. Because the random variable V_{AB} varies among the four methods, $E_{\tau_{AB}}[V_{AB}]$ must be derived separately for each of the four original estimators (GLASS, SD, STEAC, and MAC) in order to define our four improved estimators (iGLASS, iSD, iSTEAC, and iMAC). Following the notation established above for the divergence times, denote by $\breve{V}_{AB}, \bar{\breve{V}}_{AB}, \breve{\bar{V}}_{AB}$, and $\bar{\bar{V}}_{AB}$ the values of V_{AB} for the GLASS, SD, STEAC, and MAC methods, respectively.

2.1. GLASS

As derived by Jewett and Rosenberg (2012), the iGLASS estimator can be obtained by first deriving $E_{\tau_{AB}}[\breve{V}_{AB}]$ in the case of a single locus ℓ, or $E_{\tau_{AB}}[\breve{V}^{(\ell)}_{AB}]$, with $n^{(\ell)}_A$ and $n^{(\ell)}_B$ lineages sampled from species A and B, respectively. We briefly review this derivation here, as components of it are required for obtaining the other improved estimators.

2.1.1. Derivation of $E_{\tau_{AB}}[\breve{V}_{AB}]$ for a single locus. Define $K^{(\ell)}_A$ and $K^{(\ell)}_B$ to be the random numbers of lineages remaining from taxa A and B at locus ℓ at the divergence time. Conditioning on $K^{(\ell)}_A = k^{(\ell)}_A$ and $K^{(\ell)}_B = k^{(\ell)}_B$, the expectation of $\breve{V}^{(\ell)}_{AB}$ is given by Jewett and Rosenberg as

$$E_{\tau_{AB}}\left[\breve{V}^{(\ell)}_{AB}\right] = \sum_{k^{(\ell)}_A=1}^{n^{(\ell)}_A} \sum_{k^{(\ell)}_B=1}^{n^{(\ell)}_B} E\left[\breve{V}^{(\ell)}_{AB} | K^{(\ell)}_A = k^{(\ell)}_A, k^{(\ell)}_B = k^{(\ell)}_B\right] \times h_{n^{(\ell)}_A, k^{(\ell)}_A}(\tau_{AB}) \times h_{n^{(\ell)}_B, k^{(\ell)}_B}(\tau_{AB}), \tag{4}$$

where $h_{n,k}(\tau_{AB})$ is the probability that n initial lineages coalesce to k lineages in time τ_{AB}. This probability has been calculated by Tavaré (1984), and it is

$$h_{n,k}(\tau_{AB}) = \sum_{i=k}^{n} \frac{(2i-1)(-1)^{i-k} k_{(i-1)} n[i]}{k!(i-k)! n_{(i)}} exp\left[-\binom{i}{2}\tau_{AB}\right], \tag{5}$$

with $n_{[i]} = \frac{n!}{(n-i)!}$ and $n_{(i)} = \frac{(n-1+i)!}{(n-1)!}$; the time τ is in units of $2N$ generations. The conditional expectation of $\breve{V}^{(\ell)}_{AB}$, $E_{\tau_{AB}}[\breve{V}^{(\ell)}_{AB} | K^{(\ell)}_A = k^{(\ell)}_A, K^{(\ell)}_B = k^{(\ell)}_B]$, can be calculated by integrating the probability density function of $\breve{V}^{(\ell)}_{AB}$ conditional on $k^{(\ell)}_A$ and $k^{(\ell)}_B$:

$$E_{\tau_{AB}}\left[\check{V}_{AB}^{(\ell)}|K_A^{(\ell)}=k_A^{(\ell)},k_B^{(\ell)}=k_B^{(\ell)}\right]=\int_0^\infty \nu f_{\check{V}_{AB}^{(\ell)}}(\nu|k_A^{(\ell)},k_B^{(\ell)})\,d\nu. \tag{6}$$

The density $f_{\check{V}_{AB}^{(\ell)}}(\nu|k_A^{(\ell)},k_B^{(\ell)})$ is obtained by conditioning on M, the number of coalescent events that occur in the ancestral population up to and including the first interspecific coalescence; using a derivation from Takahata (1989), Jewett and Rosenberg obtained the expectation of $\check{V}_{AB}^{(\ell)}$ as a sum over all possible values of $k_A^{(\ell)}$ and $k_B^{(\ell)}$,

$$E_{\tau_{AB}}\left[\check{V}_{AB}^{(\ell)}\right]=\sum_{k_A^{(\ell)}=1}^{n_A^{(\ell)}}\sum_{k_B^{(\ell)}=1}^{n_B^{(\ell)}}\sum_{m=1}^{k^{(\ell)}-1}\Pr(M=m)\sum_{i=1}^{m}\frac{c_{i,m}}{\gamma_i}h_{n_A^{(\ell)},k_A^{(\ell)}}(\tau_{AB})h_{n_B^{(\ell)},k_B^{(\ell)}}(\tau_{AB}), \tag{7}$$

where $c_{i,\,m}=\prod_{\substack{j=1\\ j\neq i}}^{m}\frac{\gamma_j}{(\gamma_j-\gamma_i)}$ and $\gamma_i=\binom{k^{(\ell)}-(i-1)}{2}$, and where

$$\Pr(M=m)=\frac{I_{k^{(\ell)}-m,1}}{2^{m-1}k_A^{(\ell)}k_B^{(\ell)}I_{k^{(\ell)},1}}\sum_{\eta=\max\{0,\,m-k_B^{(\ell)}\}}^{\min\{m-1,\,k_A^{(\ell)}-1\}}\binom{m-1}{\eta}\left(k_{A_{[\eta+1]}}^{(\ell)}\right)^2\left(k_{B_{[m-\eta]}}^{(\ell)}\right)^2.$$

Here, $k^{(\ell)}=k_A^{(\ell)}+k_B^{(\ell)}$ and $I_{k,\,m}=\frac{k!(k-1)!}{2^{k-m}m!(m-1)!}$.

2.1.2. Derivation of $E_{\tau_{AB}}[\check{\bar{V}}_{AB}]$ for multiple loci. To derive the iGLASS correction for multiple loci, Jewett and Rosenberg (2012) obtained the conditional expectation,

$$\begin{aligned}E_{\tau_{AB}}\left[\check{\bar{V}}_{AB}|k_A^{(1)},k_B^{(1)},\ \ldots,k_A^{(L)},k_B^{(L)}\right]=&\sum_{m_1=1}^{k_1-1}\sum_{i_1=1}^{m_1}\cdots\sum_{m_L=1}^{k_L-1}\sum_{i_L=1}^{m_L}\Pr(M_1=m_1)\cdots\Pr(M_L=m_L)\\&\times c_{i_1,\,m_1}\cdots c_{i_L,\,m_L}\frac{1}{\gamma_{i_1}+\cdots+\gamma_{i_L}}.\end{aligned} \tag{8}$$

The unconditional expectation $E_{\tau_{AB}}[\check{\bar{V}}_{AB}]$ is then found using a generalization of Equation (4) that can be plugged into Equation (2) to obtain the iGLASS correction by Equation (3).

2.2. SD

2.2.1. Derivation of $E_{\tau_{AB}}[\bar{\check{V}}_{AB}]$ for a single locus. The SD estimator, $\bar{\check{V}}_{AB}$, is given by the mean over all loci of the minimum interspecific coalescence times. For a single locus, $E_{\tau_{AB}}[\bar{\check{V}}_{AB}^{(\ell)}]$ is simply equivalent to $E_{\tau_{AB}}[\check{V}_{AB}^{(\ell)}]$, the expectation of the minimum interspecific coalescence time at that locus, as derived by Jewett and Rosenberg (2012) and given in Equation (7).

2.2.2. Derivation of $E_{\tau_{AB}}[\bar{\check{V}}_{AB}^{(\ell)}]$ for multiple loci. For a set of L loci, we take the mean of $E_{\tau_{AB}}[\check{V}_{AB}^{(\ell)}]$ over the L loci to obtain the SD estimator. That is, $E_{\tau_{AB}}[\bar{\check{V}}_{AB}]=\frac{1}{L}\sum_{\ell=1}^{L}E_{\tau_{AB}}[\check{V}_{AB}^{(\ell)}]$, by the linearity of the expectation operator. Using the formula for $E_{\tau_{AB}}[\check{V}_{AB}^{(\ell)}]$ in Equation (7), we obtain

$$E_{\tau_{AB}}\left[\bar{\check{V}}_{AB}\right]=\frac{1}{L}\sum_{\ell=1}^{L}\left[\sum_{k_A^{(\ell)}=1}^{n_A^{(\ell)}}\sum_{k_B^{(\ell)}=1}^{n_B^{(\ell)}}\sum_{m=1}^{k^{(\ell)}-1}\Pr(M=m)\sum_{i=1}^{m}\frac{c_{i,\,m}}{\gamma_i}h_{n_A^{(\ell)},k_A^{(\ell)}}(\tau_{AB})h_{n_B^{(\ell)},k_B^{(\ell)}}(\tau_{AB})\right]. \tag{9}$$

Similarly to the procedure for GLASS, the expectation from (9) is then inserted into Equation (2), which is then solved for τ_{AB} to obtain the iSD estimate according to Equation (3).

2.3. MAC

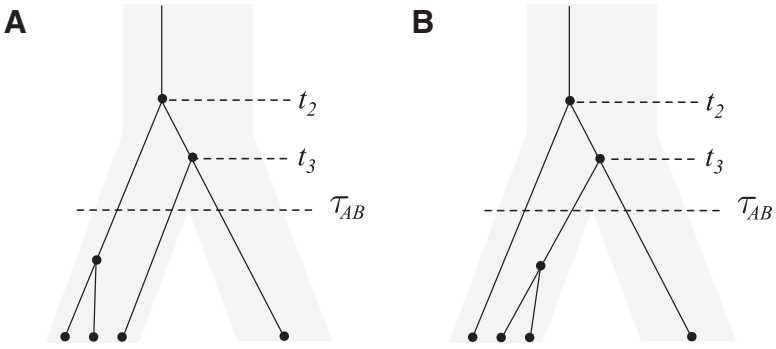
2.3.1. Derivation of $E_{\tau_{AB}}[\check{\bar{V}}_{AB}]$ for a single locus. To obtain the iMAC estimator, we begin by deriving the distribution of $\check{\bar{V}}_{AB}$ at a single locus, or $\bar{V}_{AB}^{(\ell)}$. We derive the distribution of $\bar{V}_{AB}^{(\ell)}$ by conditioning on $k_A^{(\ell)}$ and $k_B^{(\ell)}$, the numbers of lineages surviving until the divergence time. The density function of $\bar{V}_{AB}^{(\ell)}$,

however, cannot be written as simply as the analogous expression for $\check{V}_{AB}^{(\ell)}$ in the GLASS derivation. The computation of the mean pairwise interspecific coalescence time requires knowledge of the number of pairwise interspecific coalescences at each coalescent event past divergence, which in turn requires information about the topology of the gene tree under consideration.

Because many gene tree topologies are possible when more than one lineage is sampled from each taxon, we further condition on a certain set of unlabeled tree topologies, $\{Top^{(i)}\}_{i=1}^{U_{(n_A^{(\ell)}, n_B^{(\ell)}), (k_A^{(\ell)}, k_B^{(\ell)})}}$, where, as described below, $U_{(n_A^{(\ell)}, n_B^{(\ell)}), (k_A^{(\ell)}, k_B^{(\ell)})}$ is the number of unlabeled tree topologies with $(n_A^{(\ell)}, n_B^{(\ell)})$ sampled lineages and $(k_A^{(\ell)}, k_B^{(\ell)})$ lineages remaining at the divergence time. To define this set of topologies, note that given $(k_A^{(\ell)}, k_B^{(\ell)})$, the mean pairwise interspecific coalescence times—and thus, $\bar{V}_{AB}^{(\ell)}$—depend on the tree topologies both in the daughter branches of the species tree and in the ancestral branch. For example, in Figure 2, the trees in parts A and B have identical topologies past the divergence time, but the mean pairwise interspecific coalescence times differ due to different coalescence patterns in the daughter branches.

Given $(n_A^{(\ell)}, n_B^{(\ell)})$ and $(k_A^{(\ell)}, k_B^{(\ell)})$, we find the desired set of tree topologies by declaring that lineages from the same taxon are equivalent and that lineages from different taxa are distinct. Further, in the case where $n_A^{(\ell)} = n_B^{(\ell)}$, two topologies that differ only by a transposition of all of the species labels of the gene lineages are defined to be equivalent. For example, consider the case of $(n_A^{(\ell)}, n_B^{(\ell)}) = (2, 2)$. Then $k_A^{(\ell)} = 1$ or $k_A^{(\ell)} = 2$, and $k_B^{(\ell)} = 1$ or $k_B^{(\ell)} = 2$. In Table 1A, $(k_A^{(\ell)}, k_B^{(\ell)}) = (1, 1)$, and only one topology is possible; because $n_A^{(\ell)} = n_B^{(\ell)}$, the two taxa are interchangeable, and it does not matter to the computation of coalescence time distributions whether the most recent coalescence occurs in the left or the right taxon. Thus, $U_{(2, 2), (1, 1)} = 1$. Tables 1B and 1C consider $(k_A^{(\ell)}, k_B^{(\ell)}) = (2, 1)$; again, because $n_A^{(\ell)} = n_B^{(\ell)}$, it does not matter whether it is the left or the right taxon that experiences a coalescence before the divergence time, and this case is equivalent to $(k_A^{(\ell)}, k_B^{(\ell)}) = (1, 2)$. After the divergence time, the three remaining lineages form $\binom{3}{2} = 3$ pairs that are equally likely to coalesce. Because the two lineages from the right taxon are equivalent, only two distinct possibilities exist: first, as seen in Table 1B, the two lineages from the right taxon can coalesce. Alternatively, as seen in Table 1C, the lineage from the left taxon can coalesce with either of the two equivalent lineages from the right taxon. Thus, $U_{(2, 2), (2, 1)} = 2$. The four tree topologies possible when $(k_A^{(\ell)}, k_B^{(\ell)}) = (2, 2)$ appear in Tables 1D–G.

Conditional on tree topology, we calculate the density of $\bar{V}_{AB}^{(\ell)}$ by recalling from Section 2 that when n lineages are available to coalesce, the waiting time until a coalescent event is assumed under the model to be exponentially distributed with rate $\lambda = \binom{n}{2}$ coalescent time units. The mean pairwise interspecific

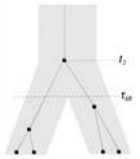
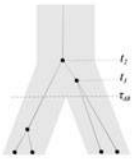
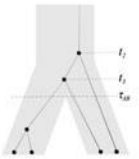
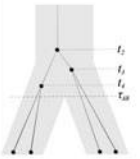
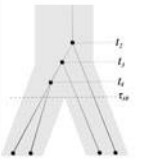
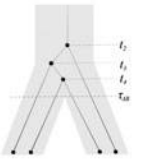
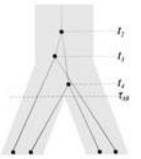


$$\bar{T} = \frac{1}{3}[(t_3) + 2(t_3 + t_2)] \qquad \bar{T} = \frac{1}{3}[2(t_3) + (t_3 + t_2)]$$

$$= t_3 + \frac{2}{3}t_2 \qquad = t_3 + \frac{1}{3}t_2$$

FIG. 2. Two trees with identical topologies above the root, but different mean interspecific coalescence times. As in Section 2.3, t_i denotes the waiting time until i lineages coalesce to $i - 1$ lineages.

TABLE 1. DISTINCT TREE TOPOLOGIES POSSIBLE FOR A LOCUS ℓ WHOSE NUMBERS OF INITIAL SAMPLED LINEAGES ARE $n_A^{(\ell)}=2$ AND $n_B^{(\ell)}=2$ AND WHOSE NUMBERS OF LINEAGES REMAINING AT DIVERGENCE ARE $k_A^{(\ell)}$ AND $k_B^{(\ell)}$

$(k_A^{(\ell)}, k_B^{(\ell)})$	(1, 1)	(2, 1) *or* (1, 2)		(2, 2)			
$Pr(k_A^{(\ell)}, k_B^{(\ell)})$	$1-2e^{-\tau_{AB}}+e^{-2\tau_{AB}}$	$2e^{-\tau_{AB}}-2e^{-2\tau_{AB}}$		$e^{-2\tau_{AB}}$			
$Top^{(i)}$							
	A	B	C	D	E	F	G
$Pr(Top^{(i)} \mid k_A^{(\ell)}, k_B^{(\ell)})$	1	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{4}{9}$	$\frac{2}{9}$
$E[\bar{V}_{AB}^{(\ell)} \mid Top^{(i)}]$	$\frac{1}{4}[4(t_2)]$ $= t_2$	$\frac{1}{4}[4(t_3+t_2)]$ $= t_3+t_2$	$\frac{1}{4}[2(t_3)+2(t_3+t_2)]$ $= t_3+\frac{1}{2}t_2$	$\frac{1}{4}[4(t_4+t_3+t_2)]$ $= t_4+t_3+t_2$	$\frac{1}{4}[2(t_4+t_3)+2(t_4+t_3+t_2)]$ $= t_4+t_3+\frac{1}{2}t_2$	$\frac{1}{4}[(t_4)+(t_4+t_3)+2(t_4+t_3+t_2)]$ $= t_4+\frac{3}{4}t_3+\frac{1}{2}t_2$	$\frac{1}{4}[(t_4)+(t_4+t_3)+2(t_4+t_3+t_2)]$ $= t_4+\frac{3}{4}t_3+\frac{1}{2}t_2$
$f_{\bar{V}_{AB}^{(\ell)}}(v \mid k_A^{(\ell)}, k_B^{(\ell)})$	e^{-v}	$\frac{1}{2}e^{-v}+4e^{-2v}-\frac{9}{2}e^{-3v}$		$\frac{2}{3}(6e^{-2v}-12e^{-4v}+6e^{-6v})+\frac{2}{9}(9e^{-2v}-12e^{-3v}+3e^{-6v})+\frac{1}{9}(\frac{9}{5}e^{-v}-3e^{-3v}+\frac{6}{5}e^{-6v})$			
$Pr(\bar{V}_{AB}^{(\ell)} \geq v)$	$=(1-2e^{-\tau_{AB}}+e^{-2\tau_{AB}})e^{-v}+(2e^{-\tau_{AB}}-2e^{-2\tau_{AB}})\left(\frac{1}{2}e^{-v}+2e^{-2v}-\frac{3}{2}e^{-3v}\right)+e^{-2\tau_{AB}}\left(\frac{2}{3}(3e^{-2v}-3e^{-4v}+e^{-6v})+\frac{2}{9}\left(\frac{9}{2}e^{-2v}-4e^{-3v}+\frac{1}{2}e^{-6v}\right)+\frac{1}{9}\left(\frac{9}{5}e^{-v}-e^{-3v}+\frac{1}{5}e^{-6v}\right)\right)$						

The set of distinct topologies, pictured in the third row, is formed by considering the $n_A^{(\ell)}$ lineages from species A to be interchangeable, but distinct from the $n_B^{(\ell)}$ interchangeable lineages from species B. Here, since $n_A^{(\ell)}=n_B^{(\ell)}$, the two taxa are also considered interchangeable. Also shown are the probabilities of each topology given $k_A^{(\ell)}$ and $k_B^{(\ell)}$ and the expected mean interspecific coalescence times conditional on each topology, denoted $E[\bar{V}^{(\ell)}|Top^{(i)}]$, which are used to derive $f_{\bar{V}_{AB}^{(\ell)}}(v|k_A^{(\ell)}, k_B^{(\ell)})$. The probabilities, $Pr(k_A^{(\ell)}, k_B^{(\ell)})$, that $k_A^{(\ell)}$ and $k_B^{(\ell)}$ lineages remain at the divergence time are also given. These probabilities are used with $f_{\bar{V}_{AB}^{(\ell)}}(v|k_A^{(\ell)}, k_B^{(\ell)})$ to calculate the unconditional density $f_{\bar{V}_{AB}^{(\ell)}}(v)$ and the survival function $Pr(\bar{V}_{AB}^{(\ell)} \geq v)$, given in the last row.

coalescence time is a weighted sum of these random variables. Consider for example the tree in Table 1A, in which all four pairwise interspecific coalescences occur at the lone interspecific coalescent event; because exactly two lineages are available to form this coalescence, we denote the waiting time until it occurs by t_2. Thus, the mean pairwise interspecific coalescence time given the topology in Table 1A is $\frac{1}{4}4t_2=t_2$. Because the mean interspecific coalescence time has the same distribution as the single interspecific coalescence time in the case that only two lineages are available to coalesce, in this case, $f_{\bar{V}_{AB}^{(\ell)}}(v|Top^{(i)}, k_A^{(\ell)}, k_B^{(\ell)})=e^{-v}$.

For a more complex example, consider Table 1F. Here, the first of the four pairwise interspecific coalescences occurs at the first coalescent event past divergence, when four lineages remain; we denote the waiting time until this coalescence by t_4. The second pairwise interspecific coalescence occurs at the next coalescent event, so that the waiting time past the divergence for this coalescent event is $t_4 + t_3$. Finally, the last two pairwise interspecific coalescences, between the rightmost lineage and each of the two lineages of the left taxon, occur at the last coalescent event, which occurs at time $t_4 + t_3 + t_2$ past the divergence time. The mean, as given in the figure, is therefore $\frac{1}{4}[(t_4)+(t_4+t_3)+2(t_4+t_3+t_2)]=t_4+\frac{3}{4}t_3+\frac{1}{2}t_2$. The distribution $f_{\bar{V}_{AB}^{(\ell)}(v|Top^{(i)}, k_A^{(\ell)}, k_B^{(\ell)})}$ can now be calculated as the sum of three exponential random variables. The first of these variables has parameter $\lambda_4=\binom{4}{2}=6$. For the second and third random variables, $\frac{3}{4}t_3$ and $\frac{1}{2}t_2$, we must consider the coefficients as well; because an exponential random variable multiplied by a constant yields a new exponential random variable, with mean given by the product of the constant and the mean of the first

variable, we multiply the parameter λ by the reciprocal of the coefficient to obtain the parameter of the new exponential random variable produced by scaling. Thus, the two remaining random variables have parameters $\lambda_3 = \binom{3}{2} \times \frac{4}{3} = 4$ and $\lambda_2 = \binom{2}{2} \times \frac{2}{1} = 2$. Because the three random variables are all exponentially distributed with distinct rates, the density of their sum can be calculated as follows:

$$f_{X_1+X_2+\cdots X_n}(x) = \left[\prod_{i=1}^{n} \lambda_i\right] \sum_{j=1}^{n} \frac{e^{-\lambda_j x}}{\prod_{\substack{k=1 \\ k\neq j}}^{n} (\lambda_k - \lambda_j)}. \tag{10}$$

In this equation (Ross, 2007), independent exponential random variables are denoted $X_1, X_2, \ldots, X_n$, with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$. Note, however, that for some topologies in cases with higher $n_A^{(\ell)}$ and $n_B^{(\ell)}$, the rate parameters will not be pairwise distinct, and a convolution integral must be used. Here, in the case of the topology in Table 1F, either method gives $f_{\bar{V}_{AB}^{(\ell)}}(v|Top^{(i)}, k_A^{(\ell)}, k_B^{(\ell)}) = 6e^{-6v} - 12e^{-4v} + 6e^{-2v}$.

Once $f_{\bar{V}_{AB}^{(\ell)}}(v|Top^{(i)}, k_A^{(\ell)}, k_B^{(\ell)})$ has been calculated for all $U_{(n_A^{(\ell)}, n_B^{(\ell)}), (k_A^{(\ell)}, k_B^{(\ell)})}$ tree topologies, we weight the density functions by conditional probabilities of the relevant topologies given $(k_A^{(\ell)}, k_B^{(\ell)})$. These probabilities, denoted $\Pr(Top^{(i)}|k_A^{(\ell)}, k_B^{(\ell)})$, are possible to compute based on the assumption of the coalescent model that each pair of lineages is equally likely to coalesce. Examining Table 1A as a simplest case, we see that if $k_A^{(\ell)} = 1$ and $k_B^{(\ell)} = 1$, only one topology is possible, and $\Pr(Top^{(1)}|k_A^{(\ell)}, k_B^{(\ell)}) = 1$. In Table 1F, however, all four lineages survive until the divergence time, so $\binom{4}{2} = 6$ choices of two lineages are possible for the first coalescent event. Four of these choices involve one lineage from each species, and therefore the probability is $\frac{2}{3}$ that the first coalescent event is compatible with the topology in Table 1F. Assuming that one of these four equivalent coalescent events occurs, $\binom{3}{2} = 3$ possible pairs of lineages are available for the next coalescent event. Because $n_A^{(\ell)} = n_B^{(\ell)}$ and the same topology will be obtained whether the lineage that coalesces with the interspecific pair is from the left or the right taxon, the probability that the next coalescence is compatible with the tree in Table 1F is also $\frac{2}{3}$. Given the first two coalescent events, the third coalescent event and the topology are determined, and so $\Pr(Top^{(3)}|k_A^{(\ell)}, k_B^{(\ell)}) = \frac{2}{3} \times \frac{2}{3} = \frac{4}{9}$. Probabilities for the other topologies are computed analogously (Table 1).

With the conditional densities of $\bar{V}_{AB}^{(\ell)}$ given a topology and the topology probabilities known, we calculate the density function for $\bar{V}_{AB}^{(\ell)}$ unconditional on tree topology,

$$f_{\bar{V}_{AB}^{(\ell)}}(v|k_A^{(\ell)}, k_B^{(\ell)}) = \sum_{i=1}^{U(n_A^{(\ell)}, n_B^{(\ell)}), (k_A^{(\ell)}, k_B^{(\ell)})} f(\bar{V}_{AB}^{(\ell)}|Top^{(i)}, k_A^{(\ell)}, k_B^{(\ell)}) \times \Pr(Top^{(i)}|k_A^{(\ell)}, k_B^{(\ell)}). \tag{11}$$

We then find the distribution of $\bar{V}_{AB}^{(\ell)}$ unconditional on $(k_A^{(\ell)}, k_B^{(\ell)})$ by summing over possible values of $(k_A^{(\ell)}, k_B^{(\ell)})$. If $n_A^{(\ell)} \neq n_B^{(\ell)}$, we sum over all $n_A^{(\ell)} \times n_B^{(\ell)}$ possible values of $k_A^{(\ell)}$ and $k_B^{(\ell)}$, as in Equation (4). When $n_A^{(\ell)} = n_B^{(\ell)} = n$, however, we avoid overcounting equivalent topologies by requiring $k_B^{(\ell)} \geq k_A^{(\ell)}$. This strategy results in a sum of $(n^2 + n)/2$ terms. Then

$$\begin{aligned} f_{\bar{V}_{AB}^{(\ell)}}(v) &= \sum_{k_A^{(\ell)}=1}^{n_A^{(\ell)}} \sum_{k_B^{(\ell)}=\alpha(n_A^{(\ell)}, n_B^{(\ell)})}^{n_B^{(\ell)}} f_{\bar{V}_{AB}^{(\ell)}}(v|k_A^{(\ell)}, k_B^{(\ell)}) \times Pr(k_A^{(\ell)}, k_B^{(\ell)}; n_A^{(\ell)}, n_B^{(\ell)}) \\ &= \sum_{k_A^{(\ell)}=1}^{n_A^{(\ell)}} \sum_{k_B^{(\ell)}=\alpha(n_A^{(\ell)}, n_B^{(\ell)})}^{n_B^{(\ell)}} f_{\bar{V}_{AB}^{(\ell)}}(v|k_A^{(\ell)}, k_B^{(\ell)}) \times h_{n_A^{(\ell)}, k_A^{(\ell)}}(\tau_{AB}) \times h_{n_B^{(\ell)}, k_B^{(\ell)}}(\tau_{AB}), \end{aligned} \tag{12}$$

where we define the starting point of the summation for k_B^{ℓ} as

$$\alpha(n_A^{(\ell)}, n_B^{(\ell)}) = \begin{cases} k_A^{(\ell)} & \text{if } n_A^{(\ell)} = n_B^{(\ell)} \\ 1 & \text{otherwise.} \end{cases}$$

2.3.2. Derivation of $E_{\tau_{AB}}[\breve{V}_{AB}]$ for multiple loci. For one locus, the probability that $\bar{V}_{AB}^{(\ell)}$ exceeds a value v is given by the survival function of $\bar{V}_{AB}^{(\ell)}$, obtained by integrating the density of $\bar{V}_{AB}^{(\ell)}$ in Equation (12) over the range $[v, \infty)$:

$$\Pr\left(\bar{V}_{AB}^{(\ell)} > v\right) = \int_{\varphi=v}^{\infty} f_{\bar{V}_{AB}^{(\ell)}}(\varphi)d\varphi. \tag{13}$$

For many loci, the survival function of $\breve{\bar{V}}_{AB}$ is given by the probability that $\bar{V}_{AB}^{(\ell)}$ exceeds v for all loci:

$$\begin{aligned}\Pr\left(\breve{\bar{V}}_{AB} > v\right) &= \Pr\left(min_{\ell}\bar{V}_{AB}^{(\ell)} > v\right) \\ &= \prod_{\ell=1}^{L}\Pr\left(\bar{V}_{AB}^{(\ell)} > v\right).\end{aligned} \tag{14}$$

In the case in which the same numbers of lineages are sampled at each locus, the survival functions for different loci are identical, and Equation (14) further simplifies to

$$\Pr\left(\breve{\bar{V}}_{AB} > v\right) = \Pr\left(\bar{V}_{AB}^{(\ell)} > v\right)^{L}. \tag{15}$$

The expectation required in computing the iMAC estimator can then be found by integration:

$$E_{\tau_{AB}}\left[\breve{\bar{V}}_{AB}\right] = \int_{v=0}^{\infty}\Pr\left(\breve{\bar{V}}_{AB} > v\right)dv. \tag{16}$$

Equation (16) gives the expectation that is inserted into Equation (2) to obtain the iMAC estimate according to Equation (3).

2.4. STEAC

The STEAC estimator of Liu et al. (2009) is found by computing the mean pairwise interspecific coalescence time for each locus, and then taking the mean of these times over all loci. Because the expected time until two lineages from different taxa coalesce is one coalescent unit past the divergence time, or $E_{\tau_{AB}}[T_{a_i^{(\ell)}, b_j^{(\ell)}}] = \tau_{AB} + 1$, the expected amount by which the STEAC estimate exceeds the divergence time is $E_{\tau_{AB}}[V_{a_i^{(\ell)}, b_j^{(\ell)}}] = 1$. The expected mean pairwise interspecific coalescence time over all pairs of lineages is therefore $E_{\tau_{AB}}[\bar{V}_{AB}^{(\ell)}] = \frac{1}{n_A^{(\ell)} n_B^{(\ell)}}\sum_{i=1}^{n_A^{(\ell)}}\sum_{j=1}^{n_B^{(\ell)}} E_{\tau_{AB}}[T_{a_i^{(\ell)}, b_j^{(\ell)}}] = 1$. Averaging across loci, $E_{\tau_{AB}}[\bar{\bar{V}}_{AB}] = \frac{1}{L}\sum_{\ell=1}^{L} E_{\tau_{AB}}[\bar{V}_{AB}^{(\ell)}] = 1$. Substituting into Equations (2) and (3), we see that the iSTEAC estimate is simply one less than the STEAC estimate when the STEAC estimate is greater than one, or zero when the STEAC estimate is less than one.

3. COMPARISON OF METHODS

3.1. Bias and mean squared error in pairwise divergence time estimates

Each of the four improved methods has the potential to substantially reduce bias in estimates of species divergence times. Some bias remains in the improved estimates, however, and we use simulations to examine bias and mean squared error (MSE) for all of the methods.

3.1.1. Simulation procedure. We performed simulations using the multispecies coalescent model, following a procedure similar to that of Jewett and Rosenberg (2012). Briefly, in each daughter branch with n sampled lineages, two lineages were chosen randomly to coalesce, with coalescence time following an exponential distribution with mean $1/\binom{n}{2}$; this procedure was repeated until only one lineage remained or until time τ, when the k_A remaining lineages from taxon A joined the k_B lineages from taxon B, and the procedure was repeated in the ancestral branch until all lineages coalesced. After obtaining a gene tree for each of a series of loci, the iGLASS, iSD, iMAC and iSTEAC estimates were calculated as in Equation (3) and Sections 2.1–2.4 from the true coalescence times taken from the set of simulated gene trees at the collection of loci. For iGLASS, the expected values from Equation (8) sometimes required excessive computation due to large numbers of nested sums. Consequently, for the improvements to GLASS, the approximate iGLASS method was used instead (Equations (20) and (21) of Jewett and Rosenberg, 2012). For simplicity, this approximation was also used for iSD.

We considered several simulated scenarios, with true divergence times set to 0, 0.1, 0.3, 0.6, 1.0, 1.5, 2.1, 2.8, and 3.6 coalescent units and with the number of loci varied from 1 to 10 in unit increments. These divergence times and numbers of loci were chosen so that trends in each parameter, as well as differences among the methods over a biologically realistic range, could potentially be detected. Simulations for each of the eight methods (GLASS, iGLASS, SD, iSD, MAC, iMAC, STEAC, and iSTEAC) were performed separately. For each method, we simulated a set of 50,000 replicate experiments for each of the 90 combinations of divergence time and number of loci, with two lineages sampled from each taxon. Two lineages were sampled for the small-sample-size simulations because with a single sampled lineage, the mean pairwise interspecific coalescence and the minimum pairwise interspecific coalescence are equivalent; differences between GLASS and MAC and between SD and STEAC would then be undetectable. To investigate sample-size effects, simulations were repeated with four lineages sampled per taxon. For these larger-sample-size simulations, the number of sampled lineages was limited to four because it is desirable to compare all methods at the same sample size, and computing the distributions of the mean pairwise interspecific coalescence times necessary for evaluation of iMAC was prohibitive with larger sample sizes.

Figure 3 shows the resulting bias, variance, and MSE in the 50,000 simulations for each of the four pairs of original and improved methods, for each of the 90 combinations of divergence time and number of loci, with two lineages sampled from each taxon. Figure 4 shows corresponding results for the case in which four lineages were sampled from each taxon.

3.1.2. Equivalences among special cases. In order to better analyze trends in the bias, variance, and MSE among the methods, we first highlight two special cases that produce equivalencies among some of our methods from shared features in the ways that the methods are constructed. First, suppose that only a single locus is sampled, as in the first rows of each heat map panel of Figure 3. Comparing the first rows of Figure 3A,C, we see that if a single locus is sampled, then GLASS and SD produce identical results. GLASS considers the minimum over loci of the locus-specific minimum coalescence times, while SD considers the mean over loci of the minimum coalescence times; for a single locus, the methods are equivalent. By the same logic, MAC, which considers the minimum over loci of locus-specific mean coalescence times, and STEAC, which considers the mean over loci of the locus-specific mean coalescence times, are also equivalent for a single locus; this result can be seen by comparing the first rows of the heat maps in Figure 3E,G. In the single-locus case, GLASS, SD, MAC, and STEAC are all identical when one lineage is sampled per taxon.

A second relationship can be seen by considering large divergence times, as shown in Figure 3 by the rightmost columns of the heat maps. As the divergence time becomes large, at any locus, the probability is high that only a single lineage will remain in each branch when the divergence time is reached. In this case, the mean and minimum interspecific coalescence at a given locus will be the same. Thus, GLASS and MAC are asymptotically equivalent at large divergence times, as are SD and STEAC. This result can be verified by comparing the rightmost columns of the heat maps in Figure 3A with those in Figure 3E, and

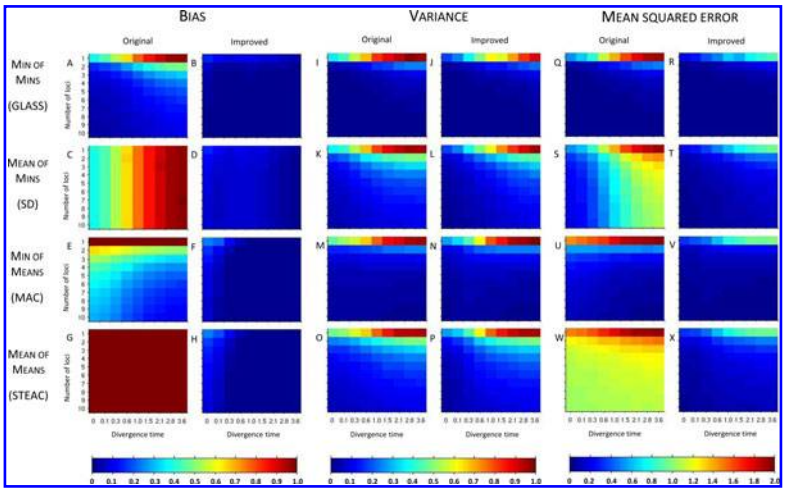


FIG. 3. Bias, variance, and mean squared error in the estimation of divergence times for the original and improved STEAC, SD, MAC, and GLASS methods. Two taxa are considered, with two lineages sampled from each taxon at each locus. Estimates were evaluated at divergence times of 0, 0.1, 0.3, 0.6, 1.0, 1.5, 2.1, 2.8, and 3.6 coalescent units and with 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 loci. 50,000 replicates were generated separately for each scenario and each method.

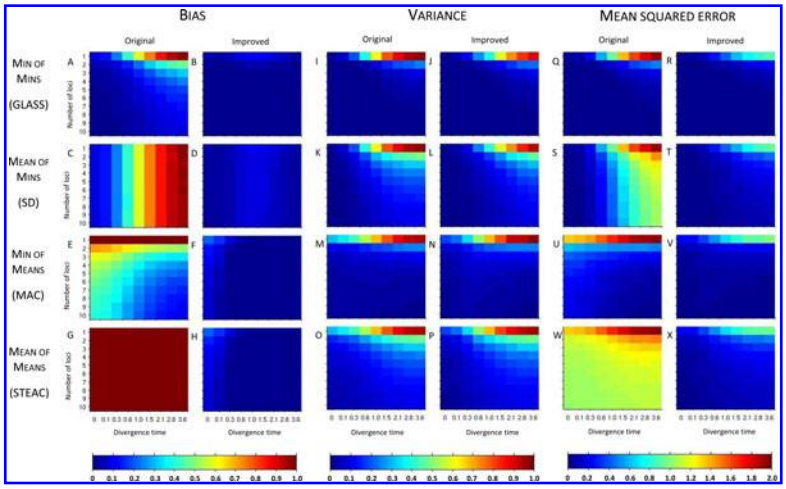


FIG. 4. Bias, variance, and mean squared error in the estimation of divergence times for the original and improved STEAC, SD, MAC, and GLASS methods, with a larger sample size than that pictured in Figure 3. Two taxa are considered, with four lineages sampled from each taxon at each locus. Estimates were evaluated at divergence times of 0, 0.1, 0.3, 0.6, 1.0, 1.5, 2.1, 2.8, and 3.6 coalescent units and with 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 loci. 50,000 replicates were generated separately for each scenario and each method.

those in Figure 3C to those in Figure 3G. At infinite divergence times in the case of a single locus, all four methods would be equivalent.

3.1.3. Bias. Considering the simulation results for the various methods individually, in Figure 3A we see that the original GLASS method has a relatively low bias compared with the other original methods (Fig. 3C,E,G). A notable improvement is still made by the approximate iGLASS method of Jewett and Rosenberg (2012), in Figure 3B, particularly when few loci are sampled and at high divergence times.

In Figure 3C, a striped pattern in the heat map for the SD method confirms that with divergence time held constant, $E_{\tau_{AB}}[\bar{\bar{V}}_{AB}]$ is constant in the number of loci. This result is a consequence of the fact that SD considers the mean over loci of the locus-specific minimum interspecific coalescence times, and when each locus has the same sample size, this mean has the same expectation irrespective of the number of loci. While improvements in bias are seen when comparing iSD, in Figure 3D, with SD, iSD retains more bias than does iGLASS.

The bias of MAC is shown in Figure 3E. As mentioned in Section 2.4, we see that for a single locus, bias is constant at a value of one coalescent unit. For more than one locus, unlike for GLASS and SD, the bias of MAC is most pronounced for short divergence times. At short divergence times, the numbers of lineages remaining at the divergence time will be larger than when the divergence time is large, and the mean pairwise interspecific coalescence time will have a smaller variance around its expected value of one coalescent unit. With few loci, the minimum of these times across loci is still likely to be quite high. As we will see in the next section, larger divergence times or more sampled loci increase the probability that one of the mean pairwise interspecific coalescence times will be small; in such cases, the MAC estimator will be less biased. Even so, for the bias in the iMAC estimator, a considerable improvement is seen over MAC (Fig. 3F).

Figure 3G shows that the bias of STEAC is constant at one coalescent unit regardless of the divergence time or number of loci sampled, as shown in Section 2.4. A marked improvement in bias for iSTEAC compared with STEAC is shown in Figure 3H, where the bias is reduced to a maximum of about 0.26 coalescent units in the case of only one locus sampled and a divergence time of zero. Note that the bias of iSTEAC is not zero because the iSTEAC estimate is set to zero rather than a negative value whenever the unimproved STEAC estimate is smaller than its smallest expected value, leading to upwardly biased estimates.

Comparing the four improved methods in Figure 3B,D,F,H, we see that all four improved methods have bias of similar magnitude. Furthermore, for each improved method, bias decreases with increasing divergence time or number of loci, with slight exceptions in the iGLASS and iSD cases resulting from the use of the approximate correction. Some differences between methods are seen for shorter divergence times or fewer loci. For small divergence times, iGLASS has the lowest bias for all of the numbers of loci that we have considered. For intermediate and long divergence times and when only a single locus is considered, iMAC and iSTEAC provide the least-biased estimates. At intermediate and long divergence times and when more than one locus is considered, iGLASS, iMAC, and iSTEAC yield the lowest biases.

3.1.4. Variance. The variances in pairwise species divergence time estimates generated by the four original and four improved methods appear in Figures 3I–P. Each unimproved method shows a pattern of variance in which values are higher at large divergence times and small numbers of loci. For larger divergence times, it is likely that only one lineage will remain in each taxon at the divergence time, and the resulting coalescence time between these two lineages will have a higher variance than will the minimum or mean coalescence time between multiple lineages. The greater variance when fewer loci are sampled is a standard effect of sampling more data: estimates based on only a few locus-specific values will have higher variance than those based on more loci.

Little change in variance is seen between any original method and its improved analogue. This result can be explained by viewing our bias reduction process as being approximately equivalent to subtracting a constant from the original estimates, an adjustment that would leave the variance unchanged. The small changes in variance that do occur arise from two deviations between the actual bias reduction method and this simplified scenario. First, in solving Equation (3), except in the case of STEAC, the amount subtracted from an estimate $\hat{t}$ to construct the improved estimate is not actually a constant, but rather, a function of $\hat{t}$ itself (Figure 5). Thus, for GLASS, SD, and MAC, the variance can change slightly between the original and improved estimates, particularly for small divergence times where the greatest deviation from the idealized scenario is seen. Second, the truncation of the improved estimates in cases in which observations lie below the expectation for zero divergence time will also decrease the variance somewhat artificially, particularly when short divergence times are considered and many improved estimates are set to zero. Because the first effect is not relevant to STEAC (Fig. 5), all of the differences between Figure 3O and Figure 3P can be attributed to the truncation effect.

The four original methods have similar variances (Fig. 3I,K,M,O), and because variance changes little between the original and improved estimates, the improved methods also have similar variances (Fig. 3J,L,N,P). As in the case of the original methods, variances for the improved methods increase with increasing divergence times and decrease with increasing numbers of loci. Variance is lowest for iGLASS, followed by iMAC, then by iSD, and iSTEAC.

3.1.5. Mean squared error. Considered together, the bias and variance results explain the patterns seen in the mean squared error (MSE) of the divergence time estimates in Figure 3Q–X. The reduction in MSE from the original to the improved methods is mostly due to the decrease in bias. Comparing the four improved methods, in Figure 3R,T,V,X, we see that the MSE in the improved methods largely shows the same pattern as the variance, increasing with greater divergence times and decreasing with greater numbers of loci. For intermediate and large divergence times, iGLASS has the lowest MSE, followed by iMAC, iSD, and iSTEAC.

3.2. Sample size

To investigate the effect of sample size on the various estimation methods, we can compare Figure 4, which considers samples of size four from each species, with Figure 3, which considers samples of size two. The comparison shows that patterns in bias, MSE, and variance are similar between the two scenarios. For large divergence times, we expect the values to be identical because at large divergence times, the probability is high that only one lineage from each branch will survive until the divergence, regardless of the initial number of lineages sampled.

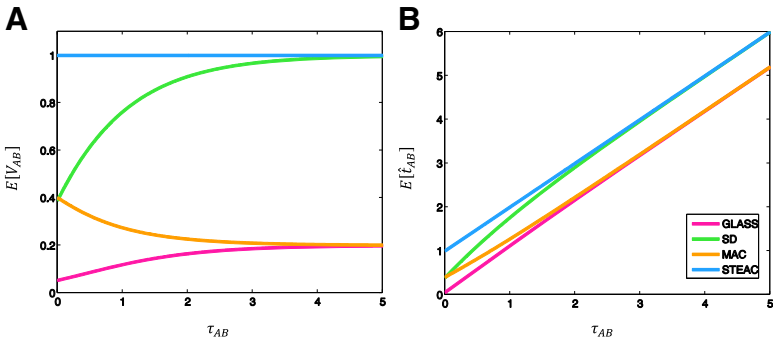


FIG. 5. $E[V_{AB}]$ and $E[\hat{t}_{AB}]$ as functions of the true divergence time τ_{AB}, for a two-taxon tree with two lineages sampled in each taxon at each of five loci. (A) Expected values of V_{AB}, the amount by which estimates exceed the true divergence time. (B) Expected values of $\hat{t}_{AB}$, the estimate of divergence time, where $E[\hat{t}_{AB}] = \tau_{AB} + E[V_{AB}]$ as in Section (2).

At small divergence times, a comparison of Figure 4A,C with Figure 3A,C shows that the original GLASS and SD methods have less bias as the number of sampled lineages increases. Because the minimum interspecific coalescent event is considered at each locus for these methods, and because we expect the first interspecific coalescent event to occur sooner after divergence when more lineages are available at the divergence time, this result is not surprising. For the MAC method in Figures 3E and 4E, particularly at small divergence times and when more than one locus is considered, we see an increase in the bias of the original method with increasing sample size. This increase occurs because the variance of the mean pairwise interspecific coalescence time at each locus decreases when more lineages are available at the divergence time to coalesce, reducing the probability that the mean pairwise interspecific coalescence time at a given locus will be near the divergence time. Figure 4G confirms that the bias in STEAC is constant with respect to sample size.

For the improved methods, slight decreases in bias are seen with increasing sample size. In Figure 3B,D,F,H, as noted above, the bias is greatest when divergence times are low. This result is due to the fact that in these cases, many estimates that lie below their expected value, and that would produce improved estimates that would underestimate the true divergence time, are set to zero by the procedure in Equation (3). The original estimates that exceed their expectations, however, still overestimate the true divergence time, and a positive bias is introduced. When sample size is increased, the variance of the original estimates decreases, as can be seen by comparing Figure 4I,K,M,O with Figure 3I,K,M,O; as a result of this decrease in variance, fewer estimates are set to zero at small divergence times, and the positive bias in the estimates is reduced. Thus, because SD experiences a relatively large reduction in variance with the increased sample size (comparing Fig. 4K with Fig. 3K), iSD shows a considerable decrease in bias with increased sample size at low divergence times (comparing Fig. 4D with Fig. 3D). MAC and STEAC, which have only slight reductions in variance with the increase in sample size, show correspondingly small reductions in the bias of their improved methods.

3.3. Trees with more than two taxa

Distance-matrix methods use pairwise distance estimates to construct a species tree through a hierarchical clustering procedure such as single-linkage clustering (Gordon, 1996) or UPGMA (Unweighted Pair Group Method with Arithmetic Mean) (Sokal and Michener, 1958). In cases with more than two species, we evaluate the accuracy in the original and improved methods by comparing divergence times of taxon pairs in the inferred tree with the corresponding divergence times in the true tree, and by comparing the topology of the inferred tree to the topology of the true tree. Note that Liu et al. (2009) performed a comparison of the unimproved GLASS, STEAC, and SD methods with respect to the inferred tree topology under a variety of scenarios; our interest here is in comparing the improved to the unimproved methods.

To evaluate the estimates of pairwise divergence times, we choose a five-taxon pectinate species tree with divergence times of 0.025, 0.02625, 0.0275, and 0.5275 coalescent units (Fig. 6); this tree, which lies

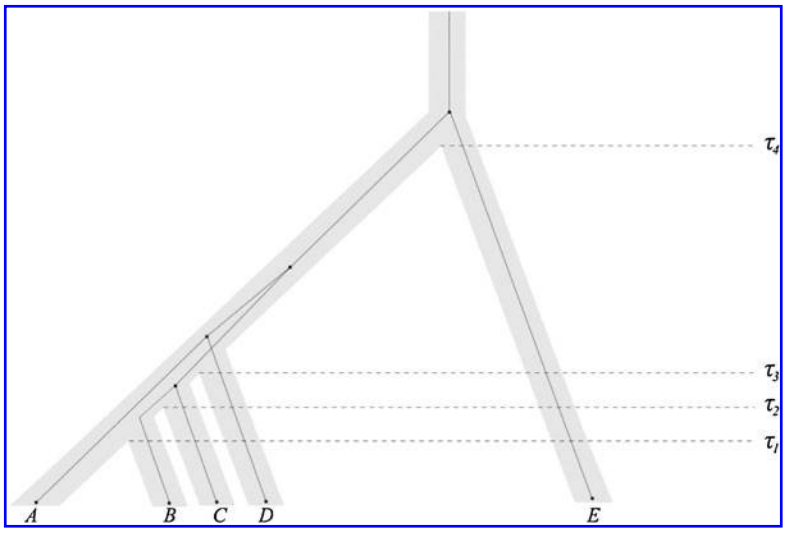


FIG. 6. The five-taxon species tree in the anomaly zone used for assessing the accuracy of estimators of pairwise divergence times. The simulation results in Figure 7 rely on this species tree. Divergence times are given by $\tau_1 = 0.025$, $\tau_2 = 0.02625$, $\tau_3 = 0.0275$, and $\tau_4 = 0.5275$ coalescent units, and the tree is not drawn to scale. Dark lines show one anomalous gene tree, (((BC)(AD))E). Under the multispecies coalescent model, this gene tree is more likely to occur than the gene tree matching the species tree, ((((AB)C)D)E).

in the anomaly zone (Degnan and Rosenberg 2006), is equivalent to a species tree used by Liu et. al (2010). Two lineages were sampled from each taxon. We first evaluate the estimation methods by disregarding the topology of the simulated tree and simply comparing divergence times of taxon pairs in the inferred tree with corresponding divergence times in the true tree. This approach allows us to obtain the bias and MSE in divergence time estimates for each pair of taxa. We performed the simulations with the coalescent simulator ms (Hudson, 2002), with 50,000 replicate sets of five gene trees representing five sampled loci. The trees generated in these 50,000 sets were used to evaluate all eight estimation methods.

To evaluate the accuracy with which each method reconstructs the topology, we consider the proportion of sampled sets of gene trees that accurately predict the species tree topology. For this analysis, a set of random five-taxon species trees was generated under the Yule model (Kulkarni, 2010). Briefly, for each tree, a split into two branches is taken at the root. Subsequent bifurcations have an equal probability of occurring on each branch, and the waiting time from the $(i-1)$st bifurcation to the (i)th bifurcation is exponentially distributed with mean $1/(i\lambda)$ For our simulations, we take $\lambda = 1$. Bifurcations continue until the external branches of a five-taxon tree have evolved, that is, until the time of the fifth bifurcation (the moment when the sixth external branch is formed). Because sampling is not likely to occur exactly at the time of the fifth bifurcation, we truncate the last branches of the species tree, with the time of truncation chosen uniformly between the times of the fourth and fifth bifurcations.

A set of 50,000 true species trees was generated following this procedure. Using the coalescent simulator ms (Hudson, 2002), one set of five gene trees, representing five loci, was then simulated for each of the true species trees, and divergence times for each pair of taxa were estimated with each of the eight methods. The same set of true species trees, and the same set of simulated gene trees for each species tree, were used to evaluate each method. Species tree estimates were constructed using a hierarchical clustering procedure. Single-linkage was defined to be the clustering method of GLASS by Mossel and Roch (2010). In single-linkage clustering, the two taxa with the shortest distance are grouped; distances between clusters are then recalculated, where the distance between two clusters A and B is defined as the shortest distance between one taxon from cluster A and one from cluster B, considering all such possible pairs of taxa. The two clusters with the shortest distance are grouped and the process is repeated until a single cluster remains. Because MAC, like GLASS, evaluates a minimum over loci, single-linkage, with its minimization procedure for computing distances between clusters, was taken to be the appropriate choice for MAC as well. To make results comparable between improved and original methods, single-linkage clustering was also used for iGLASS and iMAC. For SD, iSD, STEAC, and iSTEAC, where the mean is taken over loci, it is natural to use a clustering method that infers each node height by taking the mean over pairs of taxa that find their common ancestor at the node; because UPGMA defines the distance between cluster A and cluster

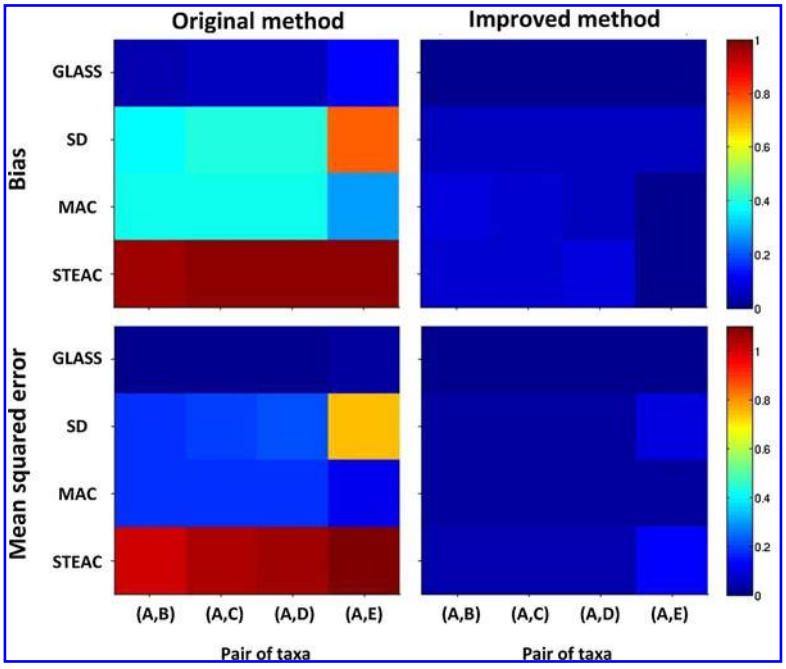


FIG. 7. Bias and mean squared error in the estimation of pairwise divergence times for a five-taxon species tree in the anomaly zone. Two lineages were sampled from each taxon at each of five loci. Values were computed using 50,000 replicate sets of five gene trees, generated separately for each method. The species tree from which gene trees were sampled is shown in Figure 6. Results are shown only for (A,B), (A,C), (A,D), and (A,E) because pairs with the same divergence times produce similar results.

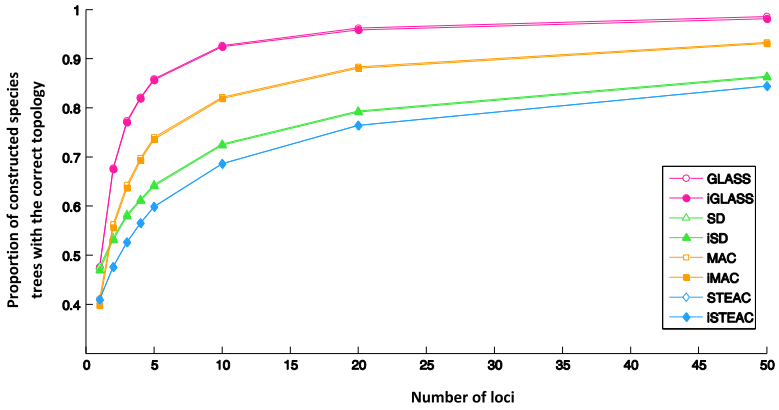


FIG. 8. Fraction of species trees reconstructed from gene trees, sampled at a given number of loci, which had the correct species tree topology. The set of species trees considered was a collection of 50,000 randomly sampled five-taxon trees, as described in Section 3.3.

B as the mean node height of all pairs of taxa, one from cluster A and the other from cluster B, UPGMA was used for these methods.

Figure 7 shows the bias and MSE observed for estimates of pairwise divergence times on the simulated data. Both bias and MSE are reduced considerably by the four improved methods, compared with the original analogues that do not incorporate a partial bias correction. The bias and MSE values are also quite similar to those in corresponding cases from Figure 3; that is, the bias and MSE in pairwise divergence time estimates in the larger tree are similar to those that would be predicted if each pair of taxa constituted a separate two-taxon tree.

Values for our second measure of accuracy, the proportion of species trees reconstructed correctly, appear in Figure 8. Each method is seen to have an increased accuracy as more loci are sampled. The differences in accuracy among the four original methods are substantial. GLASS performs best, with $>90\%$ accuracy when ten loci are considered. MAC has $>80\%$ percent accuracy at ten loci, compared with $\sim 70\%$ for SD and STEAC. Each improved method infers the correct tree topology about as often as its corresponding original method, with a slight but noticeable decrease in accuracy for iGLASS compared with GLASS and iMAC compared with MAC, and almost no difference between iSTEAC and STEAC or between iSD and SD.

Why do the original and improved methods produce similar accuracy in estimating the tree topology? For each of the methods, $g(\tau_{AB})$ is monotonic in τ_{AB} (Fig. 5B). For example, if the divergence times between taxa X and Y and between taxa Z and W were estimated to be $\hat{t}_{XY}$ and $\hat{t}_{ZW}$, respectively, with $\hat{t}_{XY} < \hat{t}_{ZW}$, then the improved estimates would also satisfy $\hat{t}^*_{XY} < \hat{t}^*_{ZW}$. Thus, none of the bias reduction procedures will change the ordering of divergence time estimates, and we would expect that the improved methods would generally construct the same species tree topology as the analogous original methods. For short divergence times, however, the truncation of estimates at zero can cause difficulties for the improved estimates. To illustrate this problem, consider a gene tree sampled from the species tree in Figure 6, with divergence time estimates obtained from any of the four original methods, such that $\hat{t}_{AB} < \hat{t}_{AC} = \hat{t}_{BC} < E_0[V_{AB}]$. Under the unimproved method, the relationships among A, B, and C would be reconstructed correctly; that is, taxa A and B would be grouped first before either was grouped with taxon C. Reducing the bias of the estimates by Equation (3), however, gives $\hat{t}^*_{AB} = \hat{t}^*_{AC} = \hat{t}^*_{BC} = 0$, and the estimated tree topology becomes ambiguous under the improved method. For the simulations in Figure 8, we dealt with this ambiguity by randomly ordering the tied estimates. For this reason, the improved methods show a slight decrease in accuracy compared with their analogous original methods. Because the original estimates must be obtained prior to the bias reduction process, a simple solution to the problem of ties would be to maintain the ordering from the original estimates in cases in which multiple estimates of zero are obtained by the improved estimator.

4. DISCUSSION

We have shown through simulations that the iGLASS, iSD, iMAC, and iSTEAC estimators of species divergence times have lower biases and mean squared errors than do the original GLASS, SD, MAC, and

STEAC estimators, and that these improvements are seen, to varying degrees, under a range of divergence times, numbers of loci, and numbers of lineages sampled. In this section, we place these results in context with respect to the literature, describe limitations of the work, and suggest some problems for future research.

As genetic studies acquire increasingly large amounts of data, computationally fast methods that reconstruct species tree topologies from gene trees while providing accurate branch length estimates are increasingly important. GLASS, SD, MAC, and STEAC provide efficient alternatives to maximum likelihood and Bayesian methods, which can be prohibitively slow on large data sets. Each of these four methods has its own distinct features. GLASS has been shown to be a consistent estimator of tree topology under the coalescent model as the number of loci increases (Mossel and Roch, 2010). SD uses the fact that minimum coalescence times between two species are consistent estimates of species divergence times as the number of sampled lineages increases (Takahata, 1989), and it is less susceptible than GLASS to erroneous inferences resulting from incorrectly inferred gene trees (Liu et al., 2009). In using means at individual loci rather than minima, MAC and STEAC are less susceptible than GLASS and SD to the possibility of divergence time estimates of zero in cases with little genetic variation; STEAC has also been shown to be consistent for the estimation of species tree topologies as the number of loci increases (Liu et al., 2009), and it is easy to compute. Our improved iSD, iMAC, and iSTEAC methods, together with the iGLASS method of Jewett and Rosenberg (2012), provide a class of estimators that can obtain more accurate branch length estimates while preserving properties that make GLASS, SD, MAC, and STEAC appealing.

Asymptotic running times for all of the original and improved methods except iMAC can be obtained by comparison to GLASS and approximate iGLASS, the complexities of which were derived by Mossel and Roch (2010) and by Jewett and Rosenberg (2012), respectively. Because taking a mean at either step in a method (either among pairwise times within a locus, or over loci, or both) instead of a minimum requires the same amount of time, GLASS, SD, MAC, and STEAC all have running time $O(n^2LS^2 + S^3)$ (Mossel and Roch, 2010), where n is the maximum number of lineages sampled from any taxon, L is the number of loci, and S is the number of taxa. iSTEAC also shares this running time, since the method entails calculating the STEAC estimate and subtracting one from each estimated divergence time. Approximate iGLASS has complexity $O(n^2LQS + LQ^3S^2)$, where Q is a tuning parameter affecting the accuracy of the computations (Jewett and Rosenberg, 2012); when the same number of lineages is sampled at each locus from each taxon, it can be shown that approximate iSD has this same precision as well. When loci differ in sample sizes, by noting that Equation C.1 of Jewett and Rosenberg (2012) must be computed L different times, it can be observed that the running time of approximate iSD becomes $O(n^2LQS + L^2Q^3S^2)$.

The computation of the iMAC estimator requires sums over all elements of a certain set of unlabeled tree topologies and over all possible values of the numbers of lineages remaining from each taxon at divergence. As the number of lineages increases, iMAC is factorial in n_A and factorial in n_B; the complexity of the calculations increases so quickly due primarily to the number of tree topologies that must be considered, but also to the increasing complexity of computing the density of the mean pairwise interspecific coalescence time conditional on each topology. We have obtained exact expressions for iMAC for cases in which up to four lineages are sampled per taxon; for large sample sizes, exact computation of the iMAC estimate is not practical, and it will be desirable to develop an approximate iMAC method similar to approximate iGLASS.

Some limitations to the applicability of our results arise from the assumptions we have made. We considered the multispecies coalescent model, in which species divergences each occur at a single point in time, with no subsequent migration or horizontal gene transfer. Further, we have assumed that the effective sizes of all populations in our model are equal, and we have not considered the effects of mutation. Perhaps most importantly, we have assumed that gene trees are known with certainty; when gene trees are inferred, errors in the inference could substantially affect the methods. A particular problem that would be encountered in real data is inferred node heights that are considerably lower than their true values. For instance, if any estimated interspecific coalescence time in the data is zero, GLASS, which takes the minimum over loci of the minimum of pairwise coalescence times, will necessarily give an estimate of zero, as will iGLASS; simulations have suggested that GLASS can have poor performance in practice (Yang and Warnow, 2011). SD, which takes the mean of minimum coalescence times, and MAC, which takes the minimum of mean coalescence times, may also be susceptible to extreme values; underestimates will then be passed on to iSD and iMAC. STEAC, on the other hand, does not employ minimization, taking a mean of mean coalescence times, and it is less likely to be affected by coalescence times of zero. STEAC has performed well for estimating species tree topologies from erroneous gene trees (Liu et al. 2009), and given a gene tree estimator that produces unbiased estimates of coalescence times, STEAC produces an unbiased mean across loci of

mean coalescence times, even if gene trees are not inferred correctly. Thus, iSTEAC, which improves upon STEAC, is also expected to perform reasonably well on estimated gene trees.

We note that all of the improved methods—iGLASS, iSD, iMAC, and iSTEAC—share a common drawback. To avoid negative estimates of speciation times, in all of the methods, we set the divergence time estimate to zero when the corresponding estimate from the original method is smaller than its smallest possible expected value, $E_0[V_{AB}]$. While this choice allows us to reduce bias without introducing unreasonable negative estimates, it can be problematic when an estimated divergence time is less than $E_0[V_{AB}]$. For example, because $E_0[V_{AB}]$ for iSTEAC is one coalescent unit, regardless of the number of loci or the numbers of lineages sampled, for any observed estimate that is below one, iSTEAC will provide an estimate of zero. This property of the method limits its utility for small divergence times. For iGLASS, iSD, and iMAC, however, this drawback is less problematic, as divergence time estimates of zero are less likely for these methods. Furthermore, for iGLASS, iSD, and iMAC, the probability of obtaining zero estimates decreases as the amount of data increases. For iGLASS and iMAC, in which a minimum is taken over loci, $E_0[V_{AB}]$ approaches zero as the number of loci increases, and for iGLASS and iSD, $E_0[V_{AB}]$ approaches zero as the number of lineages sampled at each locus in each population increases. To avoid the problem of producing estimates of zero, for all four methods, one solution is to obtain the improved estimate allowing for negative estimates, and then add the absolute value of the most negative estimate to each of the improved estimates. All estimates will then be nonnegative, and fewer estimates of zero will be produced. We expect that this approach will augment bias by an amount comparable to the bias in the most negative estimate, a value that is likely to be small.

Though our simulations have compared these four methods for a variety of values of the model parameters, a comparison of the results obtained with each of the four methods on real data or simulations with mutation would also be informative. Future work that mathematically considers the effects of mutation on these methods could improve their utility for a wider range of practical studies. Finally, an accurate approximation of iMAC would enable use of this method when large amounts of data are available.

ACKNOWLEDGMENTS

Support for this work was provided by the NSF (grants DEB-0716904 and DBI-1146722) and by the Burroughs Wellcome Fund.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

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

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

Ewing, G.B., Ebersberger, I., von Haeseler, A., et al. 2008. Rooted triple consensus and anomalous gene trees. *BMC Evol. Biol.* 8, 118.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Gordon, A.D. 1996. Hierarchical clustering, 65–121. In Arabie, P., Hubert, L.J., and De Soete, G. eds. *Clustering and Classification*. World Scientific, River Edge, NJ.

Jewett, E.M., and Rosenberg, N.A. 2012. iGLASS: An improvement to the GLASS method for estimating species trees from gene trees. *J. Comput. Biol*. 19, 293–315.

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

Kulkarni, V.G. 2010. *Modeling and Analysis of Stochastic Systems*, 2nd ed. Chapman & Hall/CRC Press, Boca Raton, FL.

Liu, L., Yu, L., and Pearl, D.K. 2010. Maximum tree: a consistent estimator of the species tree. *J. Math. Biol.* 60, 95–106.

Liu, L., Yu, L., Pearl, D.K., et al. 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., and Knowles, L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55, 21–30.

Meng, C., and Kubatko, L.S. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor. Popul. Biol.* 75, 35–45.

Mossel, E., and Roch, S. 2010. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 7, 166–171.

Ross, S.M. 2007. *Introduction to Probability Models, 9th ed.* Academic Press, New York.

Sokal, R.R., and Michener, C.D. 1958. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* 38, 1409–1438.

Takahata, N. 1989. Gene genealogy in three related populations: consistent probability between gene and population trees. Genetics 122, 957–966.

Than, C., and Nakhleh, L. 2009. Species tree inference by minimizing deep coalescences. *PLoS Comput. Biol.* 5, e1000501.

Yang, J., and Warnow, T. 2011. Fast and accurate methods for phylogenomic analyses. *BMC Bioinform.* 12, S4.

Address correspondence to:
Laura J. Helmkamp
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109

E-mail: helmkamp@umich.edu