JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 3, 2012

Pp. 293–315
DOI: 10.1089/cmb.2011.0231

iGLASS: An Improvement to the GLASS Method for Estimating Species Trees from Gene Trees

ETHAN M. JEWETT and NOAH A. ROSENBERG

ABSTRACT

Several methods have been designed to infer species trees from gene trees while taking into account gene tree/species tree discordance. Although some of these methods provide consistent species tree topology estimates under a standard model, most either do not estimate branch lengths or are computationally slow. An exception, the GLASS method of Mossel and Roch, is consistent for the species tree topology, estimates branch lengths, and is computationally fast. However, GLASS systematically overestimates divergence times, leading to biased estimates of species tree branch lengths. By assuming a multispecies coalescent model in which multiple lineages are sampled from each of two taxa at L independent loci, we derive the distribution of the waiting time until the first interspecific coalescence occurs between the two taxa, considering all loci and measuring from the divergence time. We then use the mean of this distribution to derive a correction to the GLASS estimator of pairwise divergence times. We show that our improved estimator, which we call iGLASS, consistently estimates the divergence time between a pair of taxa as the number of loci approaches infinity, and that it is an unbiased estimator of divergence times when one lineage is sampled per taxon. We also show that many commonly used clustering methods can be combined with the iGLASS estimator of pairwise divergence times to produce a consistent estimator of the species tree topology. Through simulations, we show that iGLASS can greatly reduce the bias and mean squared error in obtaining estimates of divergence times in a species tree.

Key words: algorithms, coalescence, phylogenetic trees.

1. INTRODUCTION

GENE TREES CAN DIFFER DRAMATICALLY FROM THE SPECIES TREE ON WHICH THEY EVOLVE, complicating the inference of species trees from genomic data. Discordance can arise from processes such as horizontal gene transfer and gene duplication, and in a phenomenon known as incomplete lineage sorting, it can also arise simply from randomness in the processes by which genetic lineages evolve (Maddison, 1997; Nichols, 2001; Rannala and Yang, 2008; Degnan and Rosenberg, 2009; Liu et al., 2009a). In recent years, several methods have been developed to infer species trees from gene trees, even in the presence of incomplete lineage sorting. Most of these methods, however, do not estimate branch lengths or are

Department of Biology, Stanford University, Stanford, California.

computationally slow (Maddison, 1997; Rannala and Yang, 2003; Edwards et al., 2007; Ewing et al., 2008; Degnan and Rosenberg, 2009; Kubatko et al., 2009; Liu et al., 2009b; Than and Nakhleh, 2009).

The GLASS method of Mossel and Roch (2010), which was also developed independently by Liu et al. (2010), is appealing because it estimates branch lengths, it is computationally fast, and it is a consistent estimator of the species tree topology when incomplete lineage sorting is taken to be the sole source of gene tree/species tree discordance. To estimate the species tree using the GLASS method, for each pair of taxa A and B, one first obtains an estimate $\hat{t}_{AB}$ of the divergence time τ_{AB} between A and B. The estimate $\hat{t}_{AB}$ is given by the minimum interspecific coalescence time between a lineage from taxon A and a lineage from taxon B, where the minimum is taken over all such lineage pairs and over all loci. The species tree is then constructed from the pairwise estimates by single-linkage clustering (Gordon, 1996; Mossel and Roch, 2010).

The data for the GLASS method consist of genotypes at each of L loci for a number of individuals in each taxon. Specifically, for a set of L loci indexed by $\ell = 1, \ldots, L$, let $\{a_i^\ell\}_{i=1}^{n_{A_\ell}}$ and $\{b_j^\ell\}_{j=1}^{n_{B_\ell}}$ be sets of lineages sampled at locus ℓ from taxa A and B, respectively. Let $T_{a_i^\ell, b_j^\ell}$ be an estimate of the coalescence time between lineages a_i^ℓ and b_j^ℓ, and let $T_{AB}^{(\ell)} = \min_{i,j} T_{a_i^\ell b_j^\ell}$. If τ_{AB} is the true divergence time between taxa A and B, then the GLASS estimate of τ_{AB} is given by $\hat{t}_{AB} = \min_\ell T_{AB}^{(\ell)}$, i.e., the shortest time to an interspecific coalescence at some locus (Fig. 1).

The GLASS estimate $\hat{S}$ of the species tree S is then constructed by applying single-linkage clustering to the set of estimates $\{\hat{t}_{AB}\}_{A,B\in\mathcal{S}}$, where $\mathcal{S}$ is the taxon set of the species tree. Specifically, the GLASS estimate $\hat{d}_{CC'}$ of the distance between two sets of taxa C and C' is defined by $\hat{d}_{CC'} = \min_{A\in C, B\in C'} \hat{t}_{AB}$. The single-linkage clustering procedure involves grouping the two taxon sets with shortest distance, recomputing the distances among groups, and repeating the process until a single cluster remains.

The quantity $\hat{t}_{AB}$ is a consistent estimator of the pairwise divergence time τ_{AB}, because for any $\epsilon > 0$, the probability is positive that at locus ℓ, $T_{AB}^{(\ell)}$ will exceed the divergence time by no more than ϵ time units. Thus, as more loci are sampled, it becomes increasingly likely that an interspecific coalescence at some locus will occur within ϵ time units of the divergence time τ_{AB}. The GLASS estimator $\hat{S}$ is a consistent estimator of the species tree topology, because single-linkage clustering constructs a tree with the correct topology whenever $\hat{t}_{AB}$ is close enough to τ_{AB} for all $A, B \in \mathcal{S}$.

Although the GLASS method is a consistent estimator of pairwise divergence times under the multispecies coalescent, the GLASS estimator $\hat{t}_{AB}$ systematically overestimates the divergence time τ_{AB} because interspecific coalescences occur more anciently than the divergence time under the model. It is well known that, at a given locus, the time of the first interspecific coalescence between a pair of taxa can greatly exceed the actual divergence time (Edwards and Beerli, 2000; Rosenberg and Feldman, 2002). Thus, especially when divergence times are small, the bias in GLASS estimates of divergence times can be large relative to the true times, leading to biased estimates of species tree branch lengths.

Here, by deriving the expected waiting time until the first interspecific coalescence occurs among L independent loci for a pair of taxa, we develop a correction to the GLASS estimator $\hat{t}_{AB}$. We show that the

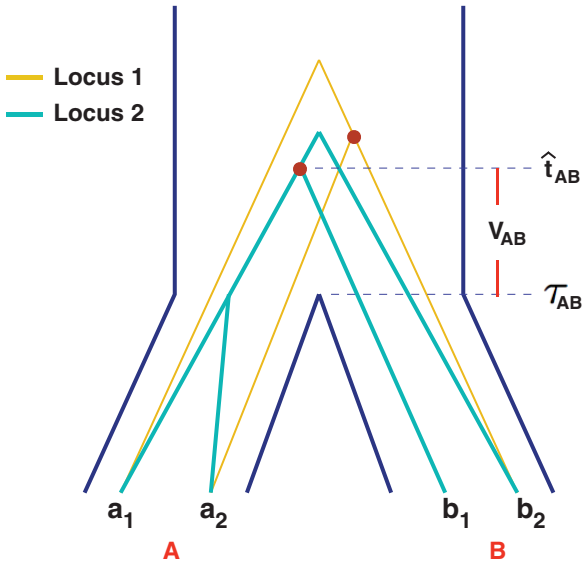

FIG. 1. The GLASS estimate of the divergence time between two taxa, A and B. Lineages a_1, a_2, b_1, and b_2 are sampled from taxa A and B, respectively, and gene trees for these lineages are shown at two loci, Locus 1 and Locus 2. Note that the individuals sampled need not be the same for all loci. The most recent interspecific coalescence at each locus is marked with a red dot. The GLASS estimate $\hat{t}_{AB}$ is the minimum interspecific coalescence time across loci. V_{AB} is the difference between the GLASS estimate and the divergence time.

corrected method, which we call iGLASS for ''improved GLASS,'' remains consistent for estimating pairwise divergence times in a species tree when incomplete lineage sorting is taken to be the sole source of gene tree discordance. We also show that each member in a particular class of clustering methods can be combined with pairwise iGLASS estimates to produce a statistically consistent estimator of the species tree topology. Through simulations, we demonstrate that in comparison with the GLASS estimator, the iGLASS estimator greatly reduces the bias and mean squared error (MSE) in pairwise estimates of species divergence times.

2. CORRECTING THE GLASS METHOD

To reduce the bias in the GLASS method's estimates of pairwise divergence times under the multispecies coalescent model, we assume that lineages evolve according to the model, and we derive the expectation of the difference $V_{AB} = \hat{t}_{AB} - \tau_{AB}$ between the GLASS estimator and the true divergence time. We then obtain a correction to the GLASS method by subtracting the expected difference $E_{\tau_{AB}}[V_{AB}]$ from the GLASS estimate $\hat{t}_{AB}$.

Under the multispecies coalescent model (Degnan and Rosenberg, 2009), in each branch of the species tree, the waiting time until i lineages coalesce to $i-1$ lineages is exponentially distributed with mean $1/\binom{i}{2}$ coalescent time units of N generations, where N is the haploid effective size of the population in the branch. All of the $\binom{i}{2}$ pairs of lineages are equally likely to coalesce. When two populations merge backwards in time, all lineages remaining in the two daughter populations enter the ancestral population, and the coalescent process resumes in that branch.

To derive the distribution of the difference V_{AB}, we model the history of each pair of species A and B using two populations with constant haploid sizes N_A and N_B. These populations merge into an ancestral population of constant size N at the divergence time τ_{AB} (Fig. 1). For simplicity, throughout this article, all times are given in units of N generations. Furthermore, although we keep our derivations general by allowing N_A and N_B to take on arbitrary values, when we consider species trees with more than two taxa, we assume that the effective population sizes are equal in every branch of the species tree, and that the species tree is binary.

At time 0, corresponding to the present, n_{A_ℓ} and n_{B_ℓ} lineages are sampled at locus ℓ ($\ell = 1, \ldots, L$) from taxa A and B, respectively. The quantities L, $\{n_{A_\ell}\}_{\ell=1}^{L}$, and $\{n_{B_\ell}\}_{\ell=1}^{L}$ are assumed to be known. We also assume that the gene trees of sampled loci have been accurately estimated. Thus, the GLASS estimate $\hat{t}_{AB}$ is exactly equal to the time of the first interspecific coalescence between taxa A and B at some sampled locus.

We assume that for each pair of taxa A and B in the species tree, each taxon in the pair has the same distance τ_{AB} (in units of N generations) from the common ancestor of A and B. This assumption implies that when times are expressed in units of N generations, the species tree that we are inferring is ultrametric. In other words, for any three taxa X, Y, and Z, two of the distances $\mathcal{D}_{XY}$, $\mathcal{D}_{XZ}$, and $\mathcal{D}_{YZ}$ are equal and are greater than or equal to the remaining distance (Semple and Steel, 2003). Ultrametricity follows from the fact that one taxon in the triplet $\{X, Y, Z\}$ is an outgroup to the other two, and we have assumed that the remaining two taxa are equidistant from it. Ultrametricity is required for the shared divergence time between a pair of taxa to be well-defined, and it also will be important for determining which clustering methods can be combined with iGLASS estimates of pairwise divergence times to produce consistent estimators of the species tree topology.

Let $\hat{t}^*_{AB}$ denote a particular value of the GLASS estimate computed from data and let $\hat{t}_{AB}$ denote the GLASS estimator, a random variable. To correct the observed GLASS estimate $\hat{t}^*_{AB}$, we find the divergence time for which the expectation of the GLASS estimator $\hat{t}_{AB}$ under the multispecies coalescent model is equal to the observed value $\hat{t}^*_{AB}$. Specifically, we solve

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

for τ_{AB}, and we take the solution as our estimate of the divergence time.

When the GLASS estimate $\hat{t}^*_{AB}$ is smaller than its smallest possible expected value $E_0[\hat{t}_{AB}]$, it is not meaningful to solve Equation (1). Therefore, we define the iGLASS estimate to be zero whenever $\hat{t}^*_{AB} < E_0[\hat{t}_{AB}]$. Defining the function $g(\tau_{AB}) = E_{\tau_{AB}}[\hat{t}_{AB}]$, our estimator $\hat{\tau}_{AB}$ of the divergence time τ_{AB}, which we call the iGLASS estimator, is given by

$$\hat{\tau}_{AB} = \begin{cases} g^{-1}(\hat{t}^*_{AB}), & \text{if } E_0[\hat{t}_{AB}] \le \hat{t}^*_{AB} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Because $E_{\tau_{AB}}[\hat{t}_{AB}]$ is a polynomial in $e^{-\tau_{AB}}$, as we will see, Equation (1) is transcendental and must be solved numerically. We now derive the quantity $E_{\tau_{AB}}[\hat{t}_{AB}]$.

3. THE EXPECTED MINIMAL INTERSPECIFIC COALESCENCE TIME $E_{\tau_{AB}}[\hat{t}_{AB}]$

Suppose that at locus ℓ ($\ell = 1, \ldots, L$), $n_{A\ell}$ and $n_{B\ell}$ lineages are sampled at time 0 from taxa A and B, respectively. Let K_{A_ℓ} and K_{B_ℓ} be random variables describing the numbers of lineages from taxa A and B remaining at the divergence time τ_{AB} at locus ℓ ($\ell = 1, \ldots, L$), and define the random vectors $\mathbf{K}_A = (K_{A_1}, \ldots, K_{A_L})$ and $\mathbf{K}_B = (K_{B_1}, \ldots, K_{B_L})$. The expectation $E_{\tau_{AB}}[\hat{t}_{AB}]$ can be expressed as $E_{\tau_{AB}}[\hat{t}_{AB}] = \tau_{AB} + E_{\tau_{AB}}[V_{AB}]$, where $V_{AB} = \hat{t}_{AB} - \tau_{AB}$ is the random difference between the GLASS estimator and the true divergence time. We now derive the expectation of V_{AB}.

Let $E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B]$ denote the expectation of V_{AB} conditional on the event that $\mathbf{K}_A = \mathbf{k}_A$ and $\mathbf{K}_B = \mathbf{k}_B$. Then

$$E_{\tau_{AB}}[V_{AB}] = \sum_{\mathbf{k}_A, \mathbf{k}_B} E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B] \prod_{\ell=1}^{L} h_{n_{A_\ell}, k_{A_\ell}}(\tau_{AB}; N_A) h_{n_{B_\ell}, k_{B_\ell}}(\tau_{AB}; N_B), \tag{3}$$

where $h_{n,k}(\tau; N_j)$ is the well-known probability that n lineages coalesce down to k lineages in time τ units of N generations in a population of constant size N_j (Tavaré, 1984). The distribution $h_{n,k}(\tau; N_j)$ is given by

$$h_{n,k}(\tau; N_j) = \sum_{i=k}^{n} \frac{(2i-1)(-1)^{i-k} k_{(i-1)} n_{[i]}}{k!(i-k)! n_{(i)}} \exp\left[-\binom{i}{2} \tau N / N_j \right], \tag{4}$$

where $n_{[i]} = \frac{n!}{(n-i)!}$ and $n_{(i)} = \frac{(n-1+i)!}{(n-1)!}$, and where the factor N/N_j comes from the fact that time is expressed in units of N generations.

The expectation $E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B]$ in Equation (3) was derived in the case of a single locus by Takahata (1989) using a recursive approach. A different recursive approach, which we present in Appendix A, can be used to compute $E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B]$ in the case of multiple loci. The desired expectation is given by

$$E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B] = \frac{1}{\sum_{\ell=1}^{L} \binom{k_{A_\ell} + k_{B_\ell}}{2}} \left[1 + \sum_{\ell=1}^{L} \left(E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A - \mathbf{e}_\ell, \mathbf{k}_B] \binom{k_{A_\ell}}{2} + E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B - \mathbf{e}_\ell] \binom{k_{B_\ell}}{2} \right) \right], \tag{5}$$

in units of N generations, where $\mathbf{e}_\ell$ is the ℓth standard basis vector of $\mathbb{R}^L$.

In addition to the mean, it is also of interest to obtain the distribution $f_{V_{AB}}(v)$ of the "overshoot" V_{AB}. Because both the unconditional probability distribution function $f_{V_{AB}}(v)$ and the conditional expectation $E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B]$ can be obtained from the conditional distribution $f_{V_{AB}}(v|\mathbf{k}_A, \mathbf{k}_B)$, we begin by computing $f_{V_{AB}}(v|\mathbf{k}_A, \mathbf{k}_B)$. We first consider the case of a single locus, and we then extend the calculation to multiple loci.

3.1. *Derivation of $f_{V_{AB}}(v|\mathbf{k}_A, \mathbf{k}_B)$ for one locus.*

Consider a single locus and let k_A and k_B denote the numbers of lineages from taxa A and B remaining at the divergence time τ_{AB}. The quantity $f_{V_{AB}}(v|k_A, k_B)$ is then the distribution of the time to the first interspecific coalescence at the locus, measuring from time τ_{AB}.

To derive $f_{V_{AB}}(v|k_A, k_B)$, recall that the time T_i until i lineages coalesce to $i-1$ lineages is exponentially distributed with mean $1/\binom{i}{2}$ Thus, if $k = k_A + k_B$ lineages remain at the divergence time, and if the first interspecific coalescence occurs on the Mth coalescence past the divergence time, then the waiting time V_{AB} until this coalescence can be expressed as the summation $V_{AB} = \sum_{i=1}^{M} T_{k-(i-1)}$.

The location M in the sequence of coalescences of the first interspecific coalescence is itself a random variable and hence, V_{AB} has a Coxian distribution (Ross, 2007) with probability density function given by

$$f_{V_{AB}}(v|k_A, k_B) = \sum_{m=1}^{k-1} Pr(M = m) \sum_{i=1}^{m} c_{i,m} \gamma_i e^{-\gamma_i v}. \tag{6}$$

In Equation (6), $c_{i,m} = \prod_{j=1, j\neq i}^{m} \gamma_j/(\gamma_j - \gamma_i)$, where $\gamma_i = \binom{k-(i-1)}{2}$ is the parameter of the ith waiting time $T_{k-(i-1)}$. For $m = 1$, we define $c_{i,m}$ to be unity.

The distribution in Equation (6) was derived by Takahata (1989). In Takahata's result, the probability $Pr(M = m)$ is obtained recursively; however, it is possible to derive a closed-form solution. Rosenberg (2003) derived a closed-form expression that is equivalent to the cumulative distribution function of M. In Appendix B, we derive the closed-form of the probability mass function of M from Equation A8 of Rosenberg (2003). We obtain

$$Pr(M=m) = \frac{I_{k-m,1}}{2^{m-1}k_A k_B I_{k,1}} \sum_{\eta=\max\{0,m-k_B\}}^{\min\{m-1,k_A-1\}} \binom{m-1}{\eta} k_{A[\eta+1]}^2 k_{B[m-\eta]}^2 \tag{7}$$

whenever $m \leq k_A + k_B - 1$, where $k_{[i]} = \frac{k!}{(k-i)!}$, and where, as in Rosenberg (2003), $I_{k,m} = \frac{k!(k-1)!}{2^{k-m}m!(m-1)!}$ is the number of ways in which k lineages can coalesce down to m lineages. Plugging expression (7) into (6) gives the formula for the distribution of V_{AB} in the case of one locus.

3.2. *Derivation of $f_{V_{AB}}(v|\mathbf{k}_A, \mathbf{k}_B)$ for L loci.*

We now extend formula (6) to multiple loci. Let $V_{AB}^{(\ell)}$ be the random variable describing the time to the first interspecific coalescence at locus ℓ ($\ell = 1, \ldots, L$). We assume that all loci are independent, conditional on the species tree and its parameter values. Therefore, the cumulative distribution function $F_{V_{AB}}(v|\mathbf{k}_A, \mathbf{k}_B)$ of the minimum interspecific coalescence time $V_{AB} = \min_\ell V_{AB}^{(\ell)}$ is given by

$$\begin{aligned} F_{V_{AB}}(v|\mathbf{k}_A, \mathbf{k}_B) &= 1 - Pr\left(V_{AB}^{(\ell)} \geq v, \forall \ell = 1, \ldots, L | \mathbf{k}_A, \mathbf{k}_B\right) \\ &= 1 - \prod_{\ell=1}^{L} Pr\left(V_{AB}^{(\ell)} \geq v | k_{A_\ell}, k_{B_\ell}\right). \end{aligned} \tag{8}$$

Here, $Pr(V_{AB}^{(\ell)} \geq v|k_{A_\ell}, k_{B_\ell})$ is given by integrating Equation (6):

$$\begin{aligned} Pr(V_{AB}^{(\ell)} \geq v|k_{A_\ell}, k_{B_\ell}) &= \int_{t=v}^{\infty} f_{V_{AB}^{(\ell)}}(t|k_{A_\ell}, k_{B_\ell})dt \\ &= \sum_{m_\ell=1}^{k_\ell-1} Pr(M_\ell = m_\ell) \sum_{i_\ell=1}^{m_\ell} c_{i_\ell,m_\ell} e^{-\gamma_{i_\ell} v}, \end{aligned} \tag{9}$$

where $k_\ell = k_{A\ell} + k_{B\ell}$ is the total number of lineages remaining at the divergence time at locus ℓ. Plugging (9) into (8) and differentiating gives the density function $f_{V_{AB}}(v|\mathbf{k}_A, \mathbf{k}_B)$:

$$\begin{aligned} f_{V_{AB}}(v|\mathbf{k}_A, \mathbf{k}_B) &= \frac{d}{dv}\left(1 - \prod_{\ell=1}^{L} Pr(V_{AB}^{(\ell)} \geq v|k_{A_\ell}, k_{B_\ell})\right) \\ &= -\sum_{\ell=1}^{L}\left[\prod_{\substack{j=1 \\ j\neq\ell}}^{L} Pr(V_{AB}^{(j)} \geq v|k_{A_j}, k_{B_j})\right] \frac{d}{dv} Pr\left(V_{AB}^{(\ell)} \geq v|k_{A_\ell}, k_{B_\ell}\right) \\ &= \sum_{\ell=1}^{L}\left[\sum_{m_1=1}^{k_1-1}\sum_{i_1=1}^{m_1} \cdots \sum_{m_L=1}^{k_L-1}\sum_{i_L=1}^{m_L} Pr(M_1 = m_1)\cdots Pr(M_L = m_L)\right. \\ &\qquad \left. \times\ c_{i_1,m_1}\cdots c_{i_L,m_L}\gamma_{i_\ell} e^{-(\gamma_{i_1}+\cdots+\gamma_{i_L})v}\right] \\ &= \sum_{m_1=1}^{k_1-1}\sum_{i_1=1}^{m_1} \cdots \sum_{m_L=1}^{k_L-1}\sum_{i_L=1}^{m_L} Pr(M_1 = m_1)\cdots Pr(M_L = m_L) \\ &\qquad \times c_{i_1,m_1}\cdots c_{i_L,m_L}(\gamma_{i_1} + \cdots + \gamma_{i_L}) e^{-(\gamma_{i_1}+\cdots+\gamma_{i_L})v}. \end{aligned} \tag{10}$$

In the last equality, we have brought the outer summation inside.

3.3. Closed-form expressions for $f_{V_{AB}}(v)$ and $E_{\tau_{AB}}[V_{AB}]$.

Closed-form expressions for $f_{V_{AB}}(v)$ and $E_{\tau_{AB}}[V_{AB}]$ can now be computed using Equation (10). The unconditional density $f_{V_{AB}}(v)$ is given by

$$f_{V_{AB}}(v) = \sum_{\mathbf{k}_A, \mathbf{k}_B} f_{V_{AB}}(v|\mathbf{k}_A, \mathbf{k}_B) \prod_{\ell=1}^{L} h_{n_{A_\ell}, k_{A_\ell}}(\tau_{AB}; N_A) h_{n_{B_\ell}, k_{B_\ell}}(\tau_{AB}; N_B), \tag{11}$$

where the summation at a given locus ℓ ($\ell = 1, \ldots, L$) in a given taxon (A or B) ranges from 1 to the number of sampled lineages at that locus in that taxon. The conditional expected value of V_{AB} for a collection of L loci is obtained by integrating Equation (10). This gives

$$\begin{aligned} E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B] &= \int_{t=0}^{\infty} t f_{V_{AB}}(t|\mathbf{k}_A, \mathbf{k}_B) dt \\ &= \sum_{m_1=1}^{k_1-1} \sum_{i_1=1}^{m_1} \cdots \sum_{m_L=1}^{k_L-1} \sum_{i_L=1}^{m_L} Pr(M_1 = m_1) \cdots Pr(M_L = m_L) \\ &\qquad \times c_{i_1, m_1} \cdots c_{i_L, m_L} \frac{1}{(\gamma_{i_1} + \cdots + \gamma_{i_L})}. \end{aligned} \tag{12}$$

The unconditional expected value $E_{\tau_{AB}}[V_{AB}]$ can be computed by plugging either Equation (12) or the recursive Equation (5) into Equation (3), thereby completing the derivation of $E_{\tau_{AB}}[V_{AB}]$.

Thus, to obtain the iGLASS estimate from the GLASS estimate $\hat{t}_{AB}$, we evaluate Equation (2), where

$$g(\tau_{AB}) = \tau_{AB} + \sum_{\mathbf{k}_A, \mathbf{k}_B} E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B] \prod_{\ell=1}^{L} h_{n_{A_\ell}, k_{A_\ell}}(\tau_{AB}; N_A) h_{n_{B_\ell}, k_{B_\ell}}(\tau_{AB}; N_B), \tag{13}$$

and where $E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B]$ is given by either Equation (12) or Equation (5). The product $\prod_{\ell=1}^{L} h_{n_{A_\ell}, k_{A_\ell}}(\tau_{AB}; N_A) h_{n_{B_\ell}, k_{B_\ell}}(\tau_{AB}; N_B)$ is a polynomial in $e^{-\tau_{AB}}$ and thus, the inverse $g^{-1}(\hat{t}_{AB})$ must be evaluated numerically. The iGLASS estimate of the species tree is then constructed by applying an appropriately chosen clustering method to the distance matrix of pairwise iGLASS time estimates. We discuss the choice of clustering method in Section 7.

4. AN APPROXIMATION

The expectation (3) is expensive to compute either when the exact formula (Equation 12) is used or when the recursion (Equation 5) is used, due to the need to sum over all possible values of k_{A_ℓ} and k_{A_ℓ}. For this reason, we introduce a deterministic approximation that amounts to an assumption that, with probability one, the number of lineages remaining at the divergence time after coalescence along a species tree branch is the number expected at that time under the coalescent model. Thus, in our approximation, Equation (3) simplifies to

$$E_{\tau_{AB}}[V_{AB}] \approx E_{\tau_{AB}}[V_{AB}|E_{\tau_{AB}}[\mathbf{K}_A], E_{\tau_{AB}}[\mathbf{K}_B]]. \tag{14}$$

Using the approximation (14) eliminates the need to sum over all possible values of k_{A_ℓ} and k_{B_ℓ}, significantly reducing the computational cost.

However, we cannot implement this approximation using our current formulas because our expression for $E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B]$, Equation (12), requires $\mathbf{k}_A$ and $\mathbf{k}_B$ to be vectors of integers, whereas $E_{\tau_{AB}}[K_{A_\ell}]$ and $E_{\tau_{AB}}[K_{B_\ell}]$ need not be integers. Although it is an option to round each expected value, $E_{\tau_{AB}}[K_{A_\ell}]$ and $E_{\tau_{AB}}[K_{B_\ell}]$ ($\ell = 1, \ldots, L$), to the nearest integer, the approximation that results is somewhat imprecise. Thus, we take a different approach and re-derive an approximation to Equation (12) in such a way that it depends continuously on the number of lineages remaining at the divergence time.

Our approach is to treat the number of lineages as a continuous quantity. We make use of a result from Maruvka et al. (2011), who demonstrated that if the initial number of lineages is large, the number of lineages remaining at time t behaves almost deterministically and is well approximated by simple

deterministic functions that approximate the expected number of lineages at time t. We wish to be as accurate as possible, however, and we therefore approximate the number of lineages at time t by the expected number of lineages at that time (Fig. 2), rather than by an approximation to the expectation.

Define $\mathcal{E}_n^t$ to be the expected number of lineages at time t units of N generations, given that n lineages exist at time $t = 0$. The expected number of lineages $\mathcal{E}_n^t$ in a population of size N_j can be computed using Equation (4), or by the following formula from Tavaré (1984):

$$\mathcal{E}_n^t = \sum_{k=1}^{n} \frac{(2k-1)n_{[k]}}{n_{(k)}} \exp\left[-\frac{k(k-1)}{2} tN/N_j\right]. \tag{15}$$

Formula (15) applies as long as the number $n = \mathcal{E}_n^0$ of lineages at time $t = 0$ is an integer. However, as it is our goal to treat lineages as a continuous quantity, we would like to allow n to be any number greater than or equal to one.

When n is not a integer, we can introduce an ''offset'' ρ such that $\mathcal{E}_{\lceil n \rceil}^{\rho} = n$. Then for any $n \geq 1$, we define the expected number of lineages φ_n^t at time t to be

$$\varphi_n^t = \begin{cases} \mathcal{E}_n^t & \text{if } n \text{ is an integer,} \\ \mathcal{E}_{\lceil n \rceil}^{\rho + t} & \text{otherwise,} \end{cases} \tag{16}$$

where ρ is found by numerically solving $\mathcal{E}_{\lceil n \rceil}^{\rho} = n$ using Equation (15). Thus, φ_n^t is a generalization of the expected number of lineages at time t to the case in which n is not integer-valued, and it allows us to treat the number of lineages as a continuous quantity. As we will see, the approximate expectation (14) computed using the approximation (16) is quite accurate even when only one or two lineages are sampled in the population.

We now use the quantity φ_n^t to derive an approximation for $E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B]$ that depends continuously on $\mathbf{k}_A$ and $\mathbf{k}_B$. We first derive an approximation to the conditional density $f_{V_{AB}}(v|\mathbf{k}_A, \mathbf{k}_B)$ in the case of a single locus, and we then generalize to many loci.

4.1. *An approximation to $f_{V_{AB}}(v|\mathbf{k}_A, \mathbf{k}_B)$ for one locus.*

As before, consider two taxa A and B. Let k_A and k_B be the numbers of lineages, not necessarily integers, that enter the ancestral population at the divergence time from taxa A and B, respectively. For the remainder of this derivation, it will simplify the notation if we measure time from a reference point at the divergence time τ_{AB}, rather than from the present. Thus, we take $\varphi_{k_A}^t$ and $\varphi_{k_B}^t$ to be the numbers of lineages remaining at time t from taxa A and B, counting from the divergence time.

Although $\varphi_{k_A}^t$ and $\varphi_{k_B}^t$ are deterministic quantities representing the expected numbers of lineages from taxa A and B, we continue to assume that the interactions between lineages are random. We assume that, in a small time interval $[t, t + \Delta t]$, a coalescent event occurs with rate $\binom{\varphi_{k_A}^t + \varphi_{k_B}^t}{2}$, given that no interspecific coalescence has occurred by time t. In addition, given that a coalescent event occurs in the interval $[t, t + \Delta t]$, we approximate the probability that it is interspecific by $2\frac{\varphi_{k_A}^t}{\varphi_{k_A}^t + \varphi_{k_B}^t}\frac{\varphi_{k_B}^t}{\varphi_{k_A}^t + \varphi_{k_B}^t - 1}$, the conditional probability that a coalescence at time t involves one lineage from taxon A and one lineage from taxon B if the numbers of

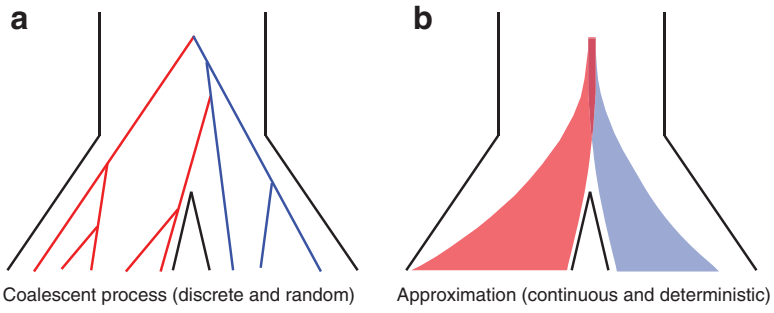

FIG. 2. Approximation to the coalescent process in a pair of populations. **(a)** A random genealogy under the standard coalescent process. **(b)** An approximation to the coalescent process in which the number of lineages at time t is the expected number of lineages. Although the number of lineages remaining from a given taxon is deterministic in our approximation, the number of interspecific coalescences that occur in some time interval Δt is random, and it depends on the approximate numbers of lineages in the two taxa.

lineages are integer-valued. Thus, letting $\mathcal{I}_{a,b}$ be the event that an interspecific coalescence occurs in the interval $[a, b]$, letting $\mathcal{I}^c_{a,b}$ be the event that an interspecific coalescence does not occur in the interval $[a, b]$, and letting $\mathcal{C}_{a,b}$ be the event that a coalescence of any kind occurs in the interval $[a, b]$, we find that

$$\begin{aligned} Pr(\mathcal{I}_{t,t+\Delta t}|\mathcal{I}^c_{0,t}) &= Pr(\mathcal{I}_{t,t+\Delta t}|\mathcal{C}_{t,t+\Delta t}, \mathcal{I}^c_{0,t}) Pr(\mathcal{C}_{t,t+\Delta t}|\mathcal{I}^c_{0,t}) \\ &\approx 2\frac{\varphi^t_{k_A}}{\varphi^t_{k_A}+\varphi^t_{k_B}} \frac{\varphi^t_{k_B}}{\varphi^t_{k_A}+\varphi^t_{k_B}-1} \binom{\varphi^t_{k_A}+\varphi^t_{k_B}}{2} \Delta t \\ &= \varphi^t_{k_A}\varphi^t_{k_B}\Delta t. \end{aligned}$$

Hence, the approximate probability that an interspecific coalescence does not occur in the interval $[t, t + \Delta t]$, given that none has occurred more recently than time t, is

$$Pr(\mathcal{I}^c_{t,t+\Delta t}|\mathcal{I}^c_{0,t}) \approx 1 - \varphi^t_{k_A}\varphi^t_{k_B}\Delta t \approx e^{-\varphi^t_{k_A}\varphi^t_{k_B}\Delta t}.$$

The probability that no interspecific coalescence occurs in the interval $[0, t]$ can be approximated by the probability that no interspecific coalescence occurs in any of J small intervals of length $\Delta t = t/J$:

$$Pr(\mathcal{I}^c_{0,t}) \approx \prod_{j=0}^{J-1} e^{-\varphi^{j\Delta t}_{k_A}\varphi^{j\Delta t}_{k_B}\Delta t} = \exp\left\{-\sum_{j=0}^{J-1}\varphi^{j\Delta t}_{k_A}\varphi^{j\Delta t}_{k_B}\Delta t\right\}.$$

Thus, as $J \to \infty$ we have $\Delta t \to 0$, and

$$Pr(\mathcal{I}^c_{0,t}) \to \exp\left\{-\int_{z=0}^{t}\varphi^z_{k_A}\varphi^z_{k_B}dz\right\}. \tag{17}$$

We now generalize this result to the case of many loci.

4.2. An approximation to $f_{V_{AB}}(v|\mathbf{k}_A, \mathbf{k}_B)$ for L loci.

Let $\varphi^t_{k_{A_\ell}}$ and $\varphi^t_{k_{A_\ell}}$ be the deterministic approximations to the numbers of lineages remaining at time t from taxa A and B at locus ℓ. Then the probability that no interspecific coalescence occurs in any one of L independent loci in the interval $[0, t]$ is approximately

$$\begin{aligned} Pr(V_{AB} \geq t|\mathbf{k}_A, \mathbf{k}_B) &\approx \prod_{\ell=1}^{L}\exp\left\{-\int_{z=0}^{t}\varphi^z_{k_{A_\ell}}\varphi^z_{k_{B_\ell}}dz\right\} \\ &= \exp\left\{-\sum_{\ell=1}^{L}\int_{z=0}^{t}\varphi^z_{k_{A_\ell}}\varphi^z_{k_{B_\ell}}dz\right\}. \end{aligned} \tag{18}$$

4.3. The approximate iGLASS correction.

To get the expected time to the first interspecific coalescence at some locus, the approximation to Equation (12), we integrate:

$$\begin{aligned} E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B] &= \int_{t=0}^{\infty} Pr(V_{AB} \geq t|\mathbf{k}_A, \mathbf{k}_B)dt \\ &\approx \int_{t=0}^{\infty}\exp\left\{-\sum_{\ell=1}^{L}\int_{z=0}^{t}\varphi^z_{k_{A_\ell}}\varphi^z_{k_{B_\ell}}dz\right\}dt. \end{aligned} \tag{19}$$

If we assume that the number of lineages remaining at the divergence time is the expected number of lineages at this time, then the approximate iGLASS correction, the approximation to Equation (3), is obtained by making the substitutions $\mathbf{k}_{A_\ell} = E_{\tau_{AB}}[\mathbf{K}_{A_\ell}]$ and $\mathbf{k}_{B_\ell} = E_{\tau_{AB}}[\mathbf{K}_{B_\ell}]$ into Equation (19):

$$\begin{aligned} E_{\tau_{AB}}[V_{AB}] &\approx E_{\tau_{AB}}[V_{AB}|E_{\tau_{AB}}[\mathbf{K}_A], E_{\tau_{AB}}[\mathbf{K}_B]] \\ &\approx \int_{t=0}^{\infty}\exp\left\{-\sum_{\ell=1}^{L}\int_{z=0}^{t}\varphi^z_{E_{\tau_{AB}}[K_{A_\ell}]}\varphi^z_{E_{\tau_{AB}}[K_{B_\ell}]}dz\right\}dt \\ &\equiv \tilde{E}_{\tau_{AB}}[V_{AB}]. \end{aligned} \tag{20}$$

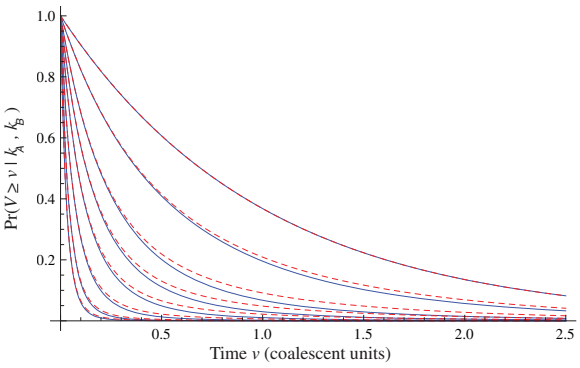

FIG. 3. Approximate survival function (Equation 18) (red, dashed) and exact survival function (Equation 9) (blue) of the quantity $V_{AB} = \hat{t}_{AB} - \tau_{AB}$ for one locus, conditional on the numbers of lineages k_A and k_B remaining at the divergence time from each taxon. $Pr(V \geq v | K_A = k_A, K_B = k_B)$ is the probability that the GLASS estimate exceeds the divergence time τ_{AB} by more than v coalescent units. In order from top to bottom, the numbers of lineages that were used to generate the curves are (k_A, k_B) = (1,1), (1,2), (2,2), (2,3), (3,3), (3,5), (5,5), (5,7), where k_A is the number of lineages remaining in taxon A at the divergence time and k_B is the corresponding number of lineages remaining in taxon B. For the top curve, one lineage is sampled from each taxon and the approximation is exact.

This approximate expression is much faster to evaluate than Equation (3) because it does not require a sum over all possible values of $k_{k_{A_\ell}}$ and $k_{k_{B_\ell}}$.

Because the values obtained from the approximation (Equation 20) differ from those obtained from the exact solution (Equation 3), we modify our definition of the iGLASS estimator (Equation 2) accordingly. We now define the function $\tilde{g}(\tau_{AB}) = \tau_{AB} + \tilde{E}_{\tau_{AB}}[V_{AB}]$, and we define the approximate iGLASS estimator $\tilde{\tau}_{AB}$ to be

$$\tilde{\tau}_{AB} = \begin{cases} \tilde{g}^{-1}(\hat{t}^*_{AB}), & \text{if } \tilde{E}_0[V_{AB}] \leq \hat{t}^*_{AB} \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

As before, the approximate iGLASS estimate $\hat{S}$ of the species tree is then constructed by applying any suitable clustering method to the pairwise approximate iGLASS estimates.

Although Equation (20) is an approximation, it can produce values that are remarkably close to the exact expectations. Figure 3 shows the exact survival function $Pr(V_{AB} \geq v)$ of V_{AB} (Equation 9) and the approximate survival function (Equation 18) for the case of one locus. From Figure 3, it can be seen that the approximation is exact when one lineage is sampled per taxon, because the expected number of lineages used in the approximation is always equal to one, the true number of lineages.

For larger numbers of sampled lineages, as the time v is increased the approximation becomes slightly worse and then improves again. This result is a consequence of the behavior of the variability in the number of lineages over time. For small v, with very high probability the number of lineages is close to the number that were initially sampled, and the variance in the number of lineages is small. For intermediate v, greater variation exists in the number of lineages, and the approximation of the stochastic process of coalescence as a deterministic process is less appropriate. Finally, for large v, the number of lineages is equal to one with high probability, and the variance is again small. Thus, the expectation φ^v_n is a better approximation to the number of lineages for small and large v.

In practice, the approximate iGLASS correction (Equation 20) differs only slightly from the exact iGLASS correction, except in the case of a single locus (Fig. 4). Therefore, in our implementation of the iGLASS correction, we use the approximation (Equation 20), except in the case of a single locus, for which it is fast to compute the exact correction.

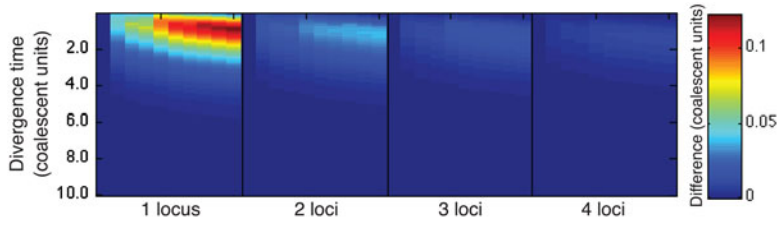

FIG. 4. The difference between the approximate iGLASS correction (Equation 19) and the exact iGLASS correction (Equation 3). Each pixel in the heatmap shows the difference $\tilde{E}_{\tau_{AB}}[V_{AB}] - E_{\tau_{AB}}[V_{AB}]$ for a given divergence time τ_{AB}, a given number of lineages sampled per taxon, and a given number of loci. Within each block corresponding to a number of loci, the numbers of lineages sampled from each taxon at each locus are, from left to right, (n_A, n_B) = (1,1), (1,2), (1,3), (1,4), (2,2), (2,3), (2,4), (3,3), (3,4), and (4,4), where n_A is the number of lineages sampled from taxon A and n_B is the number sampled from taxon B.

5. COMPUTATIONAL COMPLEXITY OF APPROXIMATE iGLASS

The computational complexity of the approximate iGLASS method is derived in Appendix C and is given by $O(n^2LQ|\mathcal{S}| + LQ^3|\mathcal{S}|^2)$ operations, where n is the maximal number of lineages sampled from any taxon at any locus, L is the number of loci, $|\mathcal{S}|$ is the number of taxa, and Q is a tuning parameter that affects the accuracy of the numerical computations (see Appendix C). For fixed Q, the estimation procedure requires at most $O(n^2L|\mathcal{S}| + L|\mathcal{S}|^2)$ operations. In comparison, the GLASS method requires $O(n^2L|\mathcal{S}|^2 + |\mathcal{S}|^3)$ operations. Thus, in each parameter, the approximate iGLASS correction has computational complexity no greater than that of GLASS for a given precision Q.

6. CONSISTENCY OF EXACT AND APPROXIMATE iGLASS

In this section, we show that both the exact and approximate iGLASS estimators (2) and (21) are consistent estimators of pairwise divergence times. We then show that applying any suitable clustering method to either exact or approximate iGLASS estimates of pairwise times produces a consistent estimator of the species tree topology. A family of clustering methods that gives rise to consistent estimation procedures is discussed in Section 7.

6.1. *Exact and approximate iGLASS are consistent estimators of pairwise divergence times.*

As we show in Theorem (D.1) in Appendix D, the GLASS method is a consistent estimator of pairwise divergence times. The exact and approximate iGLASS estimators (Equations 2 and 21) approach the GLASS estimator asymptotically in such a way that they are also consistent. We now prove this result.

Theorem 6.1. *Given two taxa, A and B, the exact iGLASS method (Equation 2) is a consistent estimator of the divergence time τ_{AB} as the number of loci $L \to \infty$.*

Proof. Let τ_{AB} be the true divergence time, and let $C_{AB} = \hat{t}_{AB} - \hat{\tau}_{AB}$ be the iGLASS correction to the GLASS method. We wish to show that $\hat{\tau}_{AB} = \hat{t}_{AB} - C_{AB}$ converges in probability to τ_{AB} as the number of loci $L \to \infty$. It is shown in Theorem D.1 that $\hat{t}_{AB} \to \tau_{AB}$ in probability as $L \to \infty$. Thus, since convergence in distribution to a constant is equivalent to convergence in probability (Casella and Berger, 2002), it follows that $\hat{t}_{AB} \to \tau_{AB}$ in distribution as $L \to \infty$. By Corollary E.3 in Appendix E, we have that $C_{AB} \leq 1/L \to 0$ as $L \to \infty$. Thus, by Slutsky's theorem (Casella and Berger, 2002), $\hat{\tau}_{AB} = \hat{t}_{AB} - C_{AB} \to \tau_{AB}$ in distribution (and in probability) as $L \to \infty$. ■

A similar result holds for the approximate iGLASS method.

Theorem 6.2. *Given two taxa, A and B, the approximate iGLASS method (Equation 21) is a consistent estimator of the divergence time τ_{AB} as the number of loci $L \to \infty$.*

Proof. In Lemma E.3 we show that the approximate iGLASS correction $\tilde{C}_{AB}$ to the GLASS estimate also satisfies $\tilde{C}_{AB} \leq 1/L$. The rest of the proof is the same as that of Theorem 6.1. ■

6.2. *Exact and approximate iGLASS are consistent estimators of the species tree topology.*

We now show that both the exact and approximate iGLASS methods are consistent estimators of the species tree topology whenever the clustering procedure applied to the estimates of pairwise divergence times has certain desirable properties. Let $\mathcal{D}$ be a distance matrix whose elements are pairwise distances between taxa in the species tree S computed according to some distance measure. Let $\hat{\mathcal{D}}$ be an estimate of $\mathcal{D}$. Let $\|A\|_\infty$ denote the magnitude of the largest element in a matrix A. Following Atteson (1999), we give the following definition.

Definition 6.3. *Let $e(S)$ denote the length of the shortest edge in a binary species tree S. Let $\mathcal{D}$ be the true matrix of pairwise distances between taxa in the tree S and let $\hat{\mathcal{D}}$ be an estimate of $\mathcal{D}$. Consider a clustering method $\mathcal{C}$ that takes a distance matrix as input and returns a tree as output. The L_∞-radius ℓ_∞ of $\mathcal{C}$ is the supremum over all quantities δ such that, for all species trees S and all estimates $\hat{\mathcal{D}}$, $\mathcal{C}$ is guaranteed to return the true topology whenever $||\hat{\mathcal{D}} - \mathcal{D}||_\infty < \delta e(S)$.*

In other words, clustering methods with nonzero L_∞-radius construct a tree with the correct topology whenever the estimated distances $\hat{\mathcal{D}}$ are close to their true values.

In our case, we are working with pairwise estimates $\{\hat{\tau}_{AB}\}_{A,B\in\mathcal{S}}$ of divergence times rather than with pairwise distances. For an ultrametric tree, the divergence time between two taxa A and B is linearly related to the distance between the taxa and is equal to half the distance in the time units in which the tree is ultrametric: in this case coalescent units, generations, or years. Thus, when the species tree S is ultrametric, the L_∞-radius of a clustering method $\mathcal{C}$ can be defined using divergence times instead of distances, as the supremum over all quantities δ such that $\mathcal{C}$ returns a tree with the correct topology whenever $\max_{A,B\in\mathcal{S}}|\hat{\tau}_{AB}-\tau_{AB}| < \delta e(S)$.

We now prove that any clustering method with nonzero L_∞-radius, when combined with a consistent estimator of pairwise divergence times, produces a consistent estimator of the species tree topology. This result was assumed by Liu et al. (2010) in their proof that GLASS is consistent. The proof is straightforward; we include it for completeness.

Proposition 6.4. *Consider a species tree S and let $\mathcal{C}$ be a clustering method with nonzero L_∞-radius ℓ_∞. Let $\hat{\tau}$ be an estimator of pairwise divergence time that is consistent as $L\to\infty$. Then the estimator $\hat{S}$ of the species tree S produced by applying clustering method $\mathcal{C}$ to the collection $\{\hat{\tau}_{AB}\}_{A,B\in\mathcal{S}}$ of divergence time estimates obtained from $\hat{\tau}$ is consistent for the tree topology as $L\to\infty$.*

Proof. Let top S denote the topology of tree S. We wish to show that $\lim_{L\to\infty} Pr(\text{top } \hat{S}=\text{top } S)=1$. We have

$$\begin{aligned} Pr(\text{top } \hat{S}=\text{top } S) &\geq Pr\left(\max_{A,B\in\mathcal{S}}|\hat{\tau}_{AB}-\tau_{AB}| < \ell_\infty e(S)\right) \\ &= 1-Pr\left(\bigcup_{A,B\in\mathcal{S}}|\hat{\tau}_{AB}-\tau_{AB}|\geq \ell_\infty e(S)\right) \\ &\geq 1-\sum_{A,B\in\mathcal{S}} Pr(|\hat{\tau}_{AB}-\tau_{AB}|\geq \ell_\infty e(S)). \end{aligned} \tag{22}$$

In the first inequality, we have used the fact that the topology of S is correctly reconstructed whenever $\max_{A,B\in\mathcal{S}}|\hat{\tau}_{AB}-\tau_{AB}| < \ell_\infty e(S)$. Since $Pr(\text{top } \hat{S}=\text{top } S)$ is a probability, we have $1-\sum_{A,B\in\mathcal{S}} Pr(|\hat{\tau}_{AB}-\tau_{AB}|\geq \ell_\infty e(S))\leq Pr(\text{top } \hat{S}=\text{top } S)\leq 1$. Since $\hat{\tau}$ is consistent, we have $\lim_{L\to\infty} Pr(|\hat{\tau}_{AB}-\tau_{AB}|\geq \ell_\infty e(S))=0$. Thus, $\lim_{L\to\infty} Pr(\text{top } \hat{S}=\text{top } S)=1$ by the ''squeeze theorem,'' proving the result. ■

It follows from results (6.1), (6.2), and (6.4) that the exact and approximate iGLASS estimators generate consistent estimators of the species tree topology when combined with any clustering method that has nonzero L_∞-radius.

7. CLUSTERING METHODS WITH NONZERO L_∞-RADIUS

Gascuel and McKenzie (2004) showed that any agglomerative algorithm defined by the following procedure (excerpted from that article) has nonzero L_∞-radius, as long as the true species tree is ultrametric:

1. Input a set of estimates of pairwise distances $\{\hat{\mathcal{D}}_{AB}\}_{A,B\in\mathcal{S}}$.
2. Choose the pair of taxa or clusters X and Y that minimize $\hat{\mathcal{D}}_{AB}$, and combine them into a new cluster U.
3. For each cluster $C\neq X, Y$, update the set of distances between C and the newly-formed cluster U according to $\hat{\mathcal{D}}_{CU}=\lambda_{UC}\hat{\mathcal{D}}_{CX}+(1-\lambda_{UC})\hat{\mathcal{D}}_{CY}$, where $\lambda_{UC}\in[0,1]$. Leave all other distances unchanged.
4. Repeat (2) and (3) until one cluster remains.

Gascuel and McKenzie (2004) reported that the class of clustering methods that follow this procedure includes single-linkage clustering (Sneath, 1957), complete-linkage clustering (Sørensen, 1948), UPGMA (Sokal and Michener, 1958), and WPGMA (Sokal and Michener, 1958). These methods differ in the choice of λ_{UC}, which is allowed to depend on U and C. For instance, Gascuel and McKenzie (2004) noted that for

single-linkage clustering, $\lambda_{UC} = 1$ when $\hat{\mathcal{D}}_{CX} \leq \hat{\mathcal{D}}_{CY}$ and $\lambda_{UC} = 0$ when $\hat{\mathcal{D}}_{CX} > \hat{\mathcal{D}}_{CY}$ (note that it is arbitrary which inequality is strict); for UPGMA, $\lambda_{UC} = |X|/(|X| + |Y|)$, where $|X|$ is the number of taxa in cluster X.

Atteson (1999) showed that the neighbor-joining method of Saitou and Nei (1987), which does not strictly follow the procedure of Gascuel and McKenzie (2004), also has nonzero L_∞-radius even when the true species tree is not ultrametric. Therefore, because we have assumed that the true species tree is ultrametric, by Proposition (6.4) we can combine neighbor-joining, or any method satisfying steps 1-4 above, with the iGLASS estimates of pairwise divergence times to produce a consistent estimator of the species tree topology.

8. A VERSION OF THE iGLASS ESTIMATOR OF PAIRWISE DIVERGENCE TIMES THAT IS UNBIASED WHEN ONE LINEAGE IS SAMPLED PER TAXON

Recall that in Equation (2), we forced the iGLASS estimates to be nonnegative. We will show that relaxing this requirement yields an unbiased estimator of pairwise divergence times in the case in which one lineage is sampled from each taxon.

Theorem 8.1. *Consider two taxa A and B. If a single lineage is sampled from each taxon at each locus* ℓ ($\ell = 1, \ldots, L$)*, then the estimator defined by* $\hat{\tau}_{AB} = g^{-1}(\hat{t}_{AB})$ *for all* $\hat{t}_{AB} \in \mathbb{R}$ *is an unbiased estimator of the divergence time* τ_{AB}.

Proof. Let k_{A_ℓ} and k_{B_ℓ} be the numbers of lineages remaining at locus ℓ ($\ell = 1, \ldots, L$) from taxa A and B at the divergence time. When one lineage is sampled from each taxon at each locus, k_{A_ℓ} and k_{B_ℓ} equal one for all $\ell = 1 \ldots, L$. Therefore, letting $\mathbf{1}$ be the vector of length L with all entries equal to 1, Equation (5) gives $E[V_{AB}|\mathbf{K}_A = \mathbf{1}, \mathbf{K}_B = \mathbf{1}] = 1/L$, and Equation (3) simplifies to $E_{\tau_{AB}}[V_{AB}] = 1/L$. The function $g(\tau_{AB})$ is then given by $g(\tau_{AB}) = \tau_{AB} + 1/L$, and its inverse by $g^{-1}(\hat{t}_{AB}) = \hat{t}_{AB} - 1/L$. Hence, $g^{-1}(t)$ is defined for all $t \in \mathbb{R}$ and it is linear. Thus, by the linearity of the expectation operator, $E_{\tau_{AB}}[g^{-1}(\hat{t}_{AB})] = g^{-1}(E_{\tau_{AB}}[\hat{t}_{AB}]) = g^{-1}(\tau_{AB} + E_{\tau_{AB}}[V_{AB}]) = g^{-1}(g(\tau_{AB})) = \tau_{AB}$. ■

This result implies that the iGLASS estimator defined by Equation (2) is also unbiased for most values of τ_{AB} whenever one lineage is sampled per taxon. Specifically, as we have assumed that gene trees are inferred with certainty, the GLASS estimate $\hat{t}_{AB}$ always exceeds the true divergence time τ_{AB}. Therefore, when one lineage is sampled per taxon at each locus and the true divergence time is greater than or equal to $1/L$, it follows that $\hat{t}_{AB} \geq 1/L$ and the iGLASS estimator is defined by $\hat{\tau}_{AB} = g^{-1}(\hat{t}_{AB}) = \hat{t}_{AB} - 1/L$. Thus, by Theorem 8.1, the iGLASS estimator will be unbiased in this case.

Note that when more than one lineage is sampled from either taxon, the probability $\prod_{\ell=1}^{L} h_{n_{A_\ell}, k_{A_\ell}}(\tau_{AB}; N_A) h_{n_{B_\ell}, k_{B_\ell}}(\tau_{AB}; N_B)$ in Equation (3) contains terms of the form $e^{-\tau_{AB}}$, and thus, the quantity $E_{\tau_{AB}}[V_{AB}]$ is no longer linear in τ_{AB}. In this case, $g^{-1}(\hat{t}_{AB})$ is not linear in $\hat{t}_{AB}$ and therefore, we cannot use the relationship $E_{\tau_{AB}}[g^{-1}(\hat{t}_{AB})] = \tau_{AB}$ when more than one lineage is sampled per taxon. However, as we will see from simulations, the bias is still very small.

9. COMPARISON OF METHODS

We used simulations to compare the performance of iGLASS to that of GLASS, evaluating each method on the basis of bias and mean squared error (MSE). We first evaluated the methods for estimating pairwise divergence times, and we then applied them to larger trees.

9.1. Simulations

We simulated gene trees under the multispecies coalescent model for various species trees S, for various numbers of loci, and for various numbers of lineages sampled per taxon. In all simulations, all population sizes were equal to the same value N across the branches of the species tree.

To simulate a gene tree from a given species tree, we used a method similar to that of Rosenberg and Feldman (2002). Let branch i refer to the branch above node i in the species tree. Let t_i be the time at node i, and let $\bar{t}_i$ be the time at the node ancestral to node i. Here, we extend our numbering to external branches, with $t_i = 0$ when i corresponds to a leaf node.

Let n_i be the number of lineages entering branch i at time t_i. If branch i is internal, then n_i is the sum of the numbers of lineages entering from its left and right daughter branches. If branch i is external, then n_i is equal to the number of lineages sampled from the corresponding taxon.

In each branch i, with the enumeration beginning with the external branches and proceeding towards the root in such a way that daughter branches have lower numbers than their parental branches, we first sampled the waiting time T_{n_i} until the first coalescence from an exponential distribution with mean $1/\binom{n_i}{2}$. If the sampled time T_{n_i} exceeded $\bar{t}_i - t_i$, then we let the set of lineages exiting branch i equal the set that entered. Otherwise, we chose two lineages at random without replacement and allowed them to coalesce. We continued in this way, at each coalescence sampling the time to the next coalescence from an exponential distribution with mean $1/\binom{q}{2}$, where q was the number of lineages remaining after the previous coalescence, until the sum of waiting times in the branch exceeded $\bar{t}_i - t_i$. The set of lineages remaining after the last coalescence to occur within branch i was then merged into the set of lineages entering its ancestral branch, along with the set of lineages entering from its sister branch, and the process was repeated in the ancestral branch. Simulations were run until all lineages coalesced to a single lineage. For trees with more than two taxa, the simulations were carried out using the software program *ms* (Hudson, 2002).

Let n_{X_ℓ} denote the number of lineages sampled from taxon X at locus ℓ ($\ell = 1, \ldots, L$). For a given species tree S together with a set of parameters consisting of a number of loci L and numbers of lineages $\{n_{X_\ell}\}_{X \in \mathcal{S}}$, we first sampled r independent sets $\mathcal{L}_j$ of L gene trees ($j = 1, \ldots, r$). For each set $\mathcal{L}_j$, we computed the GLASS estimate $\hat{t}_{AB}^{(j)}$ for all pairs of species $A, B \in \mathcal{S}$ using the GLASS algorithm (Section 1), without applying the single-linkage clustering step. From each observation $\hat{t}_{AB}^{(j)}$, we then computed an observation $\hat{\tau}_{AB}^{(j)}$ of the exact iGLASS estimate, and an observation $\tilde{\tau}_{AB}^{(j)}$ of the approximate iGLASS estimate. We thus obtained the sets of pairwise estimates $\{\hat{t}_{AB}^{(j)}\}_{A,B \in \mathcal{S}}$, $\{\hat{\tau}_{AB}^{(j)}\}_{A,B \in \mathcal{S}}$, and $\{\tilde{\tau}_{AB}^{(j)}\}_{A,B \in \mathcal{S}}$, for each set of gene trees $\mathcal{L}_j$ ($j = 1, \ldots, r$). For species trees with more than two taxa, only $\{\hat{t}_{AB}^{(j)}\}_{A,B \in \mathcal{S}}$ and $\{\tilde{\tau}_{AB}^{(j)}\}_{A,B \in \mathcal{S}}$ were computed.

For each set $\{\hat{t}_{AB}^{(j)}\}_{A,B \in \mathcal{S}}$ ($j = 1, \ldots, r$), we computed the GLASS estimate $\hat{S}^{(j)}$ of the species tree by single-linkage clustering, and for each internal node i in this estimated species tree, we estimated the height $\hat{t}_i^{(j)}$ of the node i by the distance between the two clusters combined on the step of the clustering method that produced the node. The clustering procedure was omitted for trees with two taxa because the estimates $\hat{t}_{AB}^{(j)}$ ($j = 1, \ldots, r$) already provide estimates of the divergence time τ_{AB}. We then compared each estimated node height $\hat{t}_i^{(j)}$ to its true value t_i, and we computed the average difference $\mathcal{B}^{(j)}(\hat{t}) = \sum_{i=1}^{|\mathcal{S}|-1} (\hat{t}_i^{(j)} - t_i)/(|\mathcal{S}| - 1)$ and the average squared difference $\mathcal{M}^{(j)}(\hat{t}) = \sum_{i=1}^{|\mathcal{S}|-1} (\hat{t}_i^{(j)} - t_i)^2/(|\mathcal{S}| - 1)$.

Average bias in the GLASS method was estimated by $\widehat{bias}_{avg}(\hat{t}) = \frac{1}{r}\sum_{j=1}^{r} \mathcal{B}^{(j)}(\hat{t})$, and average MSE by $\widehat{MSE}_{avg}(\hat{t}) = \frac{1}{r}\sum_{j=1}^{r} \mathcal{M}^{(j)}(\hat{t})$. The average bias and MSE in the exact and approximate iGLASS methods were estimated by the same procedure (using single-linkage clustering), but using the times $\{\hat{\tau}_{AB}^{(j)}\}_{A,B \in \mathcal{S}; j = 1, \ldots, r}$ and $\{\tilde{\tau}_{AB}^{(j)}\}_{A,B \in \mathcal{S}; j = 1, \ldots, r}$.

We denote the average bias and MSE in the exact iGLASS method by $\widehat{bias}_{avg}(\hat{\tau})$ and $\widehat{MSE}_{avg}(\hat{\tau})$, and we denote the average bias and MSE in the approximate iGLASS method by $\widehat{bias}_{avg}(\tilde{\tau})$ and $\widehat{MSE}_{avg}(\tilde{\tau})$.

9.2. Estimating pairwise divergence times.

To evaluate the performance of the three methods for estimating pairwise divergence times, we simulated gene trees under the multispecies coalescent from a species tree with two taxa, for various values of the parameters τ_{AB}, L, $\{n_{A_\ell}\}_{\ell=1}^{L}$, and $\{n_{B_\ell}\}_{\ell=1}^{L}$, and for r = 50,000 replicates. In varying the parameters $\{n_{A_\ell}\}_{\ell=1}^{L}$ and $\{n_{B_\ell}\}_{\ell=1}^{L}$, we maintained the relationships $n_{A_\ell} = n_A$ and $n_{B_\ell} = n_B$ for all ℓ.

We considered values of 1, 5, 10, and 50 for L. However, because the exact iGLASS estimate is difficult to compute in the case of both multiple loci and large numbers of lineages, only the GLASS estimate and approximate iGLASS estimate were computed when both the number of loci and the number of lineages were large.

9.2.1. Bias. Figure 5 indicates that especially for small divergence times, the bias in the GLASS estimate can be large relative to the divergence time. Whenever a single lineage is sampled from each taxon, the bias in the GLASS method is $1/L$ in coalescent units of N generations, regardless of the divergence time. One lineage always remains at the divergence time from each taxon at each locus, and therefore, the expected time to the first interspecific coalescence is the expectation of the minimum of L independent exponentially distributed random variables, each with a mean of one coalescent time unit. For

example, in a haploid population with an effective size of $N = 10,000$, if the GLASS estimate is based on a single lineage sampled from each population at each of 20 loci, then the bias in the GLASS estimate is 10,000/20 = 500 generations.

Although sampling multiple lineages from each population can greatly reduce the bias for low divergence times, it does not reduce the bias for larger divergence times. As noted by Mossel and Roch (2010), when τ_{AB} is measured in units of generations, the probability that a single lineage remains at the top of the branch corresponding to taxon A is bounded below by $1 - 3e^{-\tau_{AB}/N_A}$ and the probability that a single lineage remains at the top of the branch corresponding to taxon B is bounded below by $1 - 3e^{-\tau_{AB}/N_B}$ (Tavaré, 1984). This bound can be made arbitrarily close to one by increasing the divergence time and, as the divergence time increases, the GLASS estimate approaches the value of the GLASS estimate when one lineage is sampled per taxon, or $1/L$ coalescent units.

To compare the estimated bias in the exact and approximate iGLASS methods to the estimated bias in the GLASS method, we computed the ratios $\widehat{bias}(\hat{\tau}_{AB})/\widehat{bias}(\hat{t}_{AB})$ and $\widehat{bias}(\tilde{\tau}_{AB})/\widehat{bias}(\hat{t}_{AB})$ (Fig. 5). For most values of the divergence time, the bias in the approximate iGLASS method is negligible compared to the bias in the GLASS method; although it is considerably larger in magnitude for small values of τ_{AB}, the bias ratio continues to be less than 1. The bias is not entirely negligible in this case because we define the exact and approximate iGLASS estimates to be zero whenever the GLASS estimate is lower than its smallest possible expected time (Equations 2 and 21). Thus, when the GLASS estimate is small, instead of subtracting a positive quantity from the GLASS estimate to produce the iGLASS estimate, we estimate the divergence time to be zero, resulting in an iGLASS estimate (exact or approximate) that is biased upwards. This truncation prevents the iGLASS estimators from completely eliminating the bias, but it also leads to a decrease in variance, which ultimately leads to a lower mean squared error at these divergence times. The decrease in MSE due to lower variance can be seen by the yellow bars across the tops of the MSE graphs in Figure 5.

9.2.2. Mean squared error. The ratios $\widehat{MSE}(\hat{\tau}_{AB})/\widehat{MSE}(\hat{t}_{AB})$ and $\widehat{MSE}(\tilde{\tau}_{AB})/\widehat{MSE}(\hat{t}_{AB})$ are shown in Figure 5 for various values of τ_{AB}, n_A, and n_B. From these plots, we can see that $\widehat{MSE}(\hat{\tau}_{AB})/\widehat{MSE}(\hat{t}_{AB})$ and $\widehat{MSE}(\tilde{\tau}_{AB})/\widehat{MSE}(\hat{t}_{AB})$ are roughly 1/2, and that they appear to approach 1/2 as τ_{AB} increases.

To see why this is reasonable, consider the case in which a single lineage is sampled per taxon at each locus. In this case, the ''overshoot'' in the GLASS estimate, $V_{AB} = \hat{t}_{AB} - \tau_{AB}$, is distributed exponentially with mean $1/L$. Thus, the bias in the GLASS estimator is $E_{\tau_{AB}}[V_{AB}] = 1/L$, its variance is $\mathrm{Var}(V_{AB}) = 1/L^2$, and its MSE is $\mathrm{MSE}(\hat{t}_{AB}) = 2/L^2$. The variance in the GLASS estimator then accounts for half of the mean squared error when one lineage is sampled per taxon.

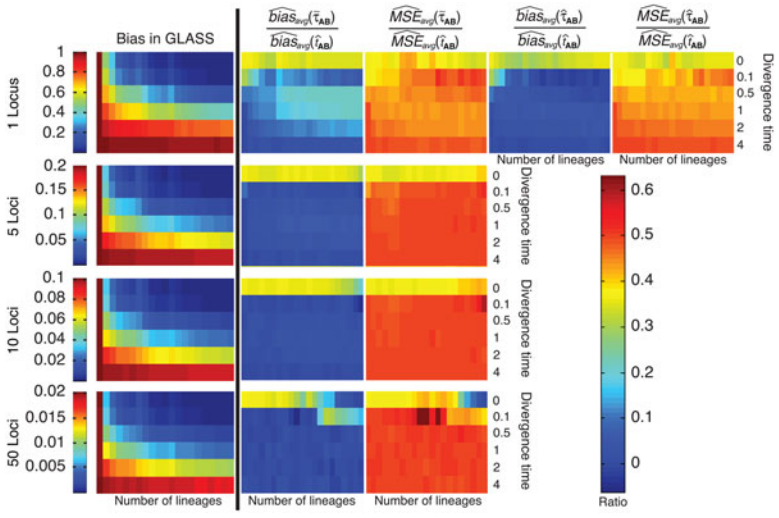

FIG. 5. Comparison of bias and mean squared error for the GLASS, exact iGLASS, and approximate iGLASS methods for two taxa and one locus. All values were computed using 50,000 simulation replicates. In each of the fourteen small heatmap panels, the divergence time between two taxa A and B is given in coalescent units on the y-axis. In each heatmap, the divergence times are, from top to bottom, $\tau_{AB} = 0$, 0.1, 0.5, 1, 2, and 4 coalescent units. In each heatmap, the numbers of lineages sampled from each taxon are given on the x-axis in the format (n_A, n_B), where n_A is the number of lineages sampled from taxon A, and n_B is the number sampled from taxon B. From left to right, the numbers of lineages in each column are (n_A, n_B) = (1,1), (1,3), (1,5), (1,10), (1,15), (1,20), (3,3), (3,5), (3,10), (3,15), (3,20), (5,5), (5,10), (5,15), (5,20), (10,10), (10,15), (10,20), (15,15), (15,20), (20,20).

When one lineage is sampled per taxon, the iGLASS correction to the GLASS estimator is computed by subtracting a constant quantity $1/L$ from the GLASS estimate, except when $\hat{t}_{AB}$ is in the region $\in [0, 1/L)$, which decreases in size as $L \to \infty$. Thus, the variance of the (exact or approximate) iGLASS estimator is nearly equal to the variance of the GLASS estimator. As Theorem 8.1 indicates, when a single lineage is sampled per taxon, the iGLASS estimator is almost unbiased. Thus, when a single lineage is sampled per taxon, the MSE in the (exact or approximate) iGLASS estimator is approximately equal to the variance in the GLASS estimator, which is half the MSE in the GLASS estimator. Because k_{A_ℓ} and k_{B_ℓ} approach one in probability as $\tau_{AB} \to \infty$, we expect that $\text{MSE}(\tilde{\tau}_{AB})$ will approach $\text{MSE}(\hat{t}_{AB})/2$ as τ_{AB} increases to infinity.

9.3. *Exact versus approximate iGLASS*

In the majority of our simulations, we have used the approximate iGLASS correction rather than the exact method because the exact correction is difficult to compute. However, consider the panels in the first row of Figure 5 that correspond to the case of one locus. It can be seen that the bias and MSE in the approximate iGLASS method are very similar to the bias and MSE in the exact iGLASS method. This result indicates that making the approximation $E_{\tau_{AB}}[V_{AB}] \approx \tilde{E}_{\tau_{AB}}[V_{AB}]$ (Equation 20) has little effect on the performance of the iGLASS estimator in the case of one locus. Because Figure 4 indicates that the approximation is least accurate in the case of a single locus, the similarity of the bias and MSE for the exact and approximate methods in the case of one locus suggests that making the approximation $E_{\tau_{AB}}[V_{AB}] \approx \tilde{E}_{\tau_{AB}}[V_{AB}]$ generally has little effect on the performance of the iGLASS method relative to that of GLASS.

9.4. *iGLASS for larger trees*

Figure 6 shows the ratios $\widehat{bias}_{avg}(\tilde{\tau}_{AB})/\widehat{bias}_{avg}(\hat{t}_{AB})$ and $\widehat{MSE}_{avg}(\tilde{\tau}_{AB})/\widehat{MSE}_{avg}(\hat{t}_{AB})$ computed over $r = 50{,}000$ replicates for two different five-taxon species trees similar to those used by Liu et al. (2010) to evaluate the performance of the GLASS method. One internal branch of the tree is short enough that the most likely gene tree given the species tree does not have the topology of the true tree. In other words, the tree is in the anomaly zone of Degnan and Rosenberg (2006).

From Figure 6, we see that the average bias in the iGLASS estimate is often considerably less than that of the GLASS estimate. The improvement in the bias is best for small numbers of loci and decreases as the number of loci increases. However, the bias in the GLASS method itself decreases quickly as the number of loci is increased.

Note that although the iGLASS correction improves the bias and MSE in the estimates of species tree node heights, it does not improve the accuracy in estimating topologies. For both species trees ((((**E**:0.5,

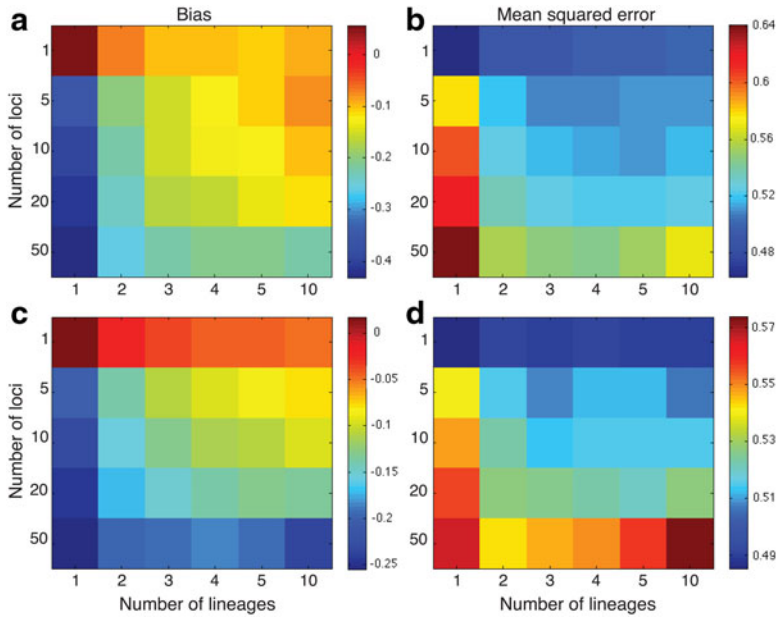

FIG. 6. Comparison of mean squared error and bias in the approximate iGLASS and GLASS methods for two five-taxon species trees used in Liu et al. (2010) to evaluate the GLASS method. In Newick format, the tree in **(a)** and **(b)**, a caterpillar, is given by ((((**E**:0.5, **D**:0.5):0.025, **C**:0.525):0.025, **B**:0.55):10.0, **A**:10.55). The tree in **(c)** and **(d)**, another 5-taxon caterpillar, is ((((**E**:0.5, **D**:0.5):0.2, **C**:0.7):1, **B**:1.7):10.0, **A**:11.7). The first tree is in the anomaly zone; the second tree is not. All values were computed using 50,000 replicates. The clustering method applied to the approximate iGLASS estimates was single-linkage. (a) and (c): The ratio $\widehat{bias}_{avg}(\tilde{\tau}_{AB})/\widehat{bias}_{avg}(\hat{t}_{AB})$. **(b)** and **(d)**: The ratio $\widehat{MSE}_{avg}(\tilde{\tau}_{AB})/\widehat{MSE}_{avg}(\hat{t}_{AB})$.

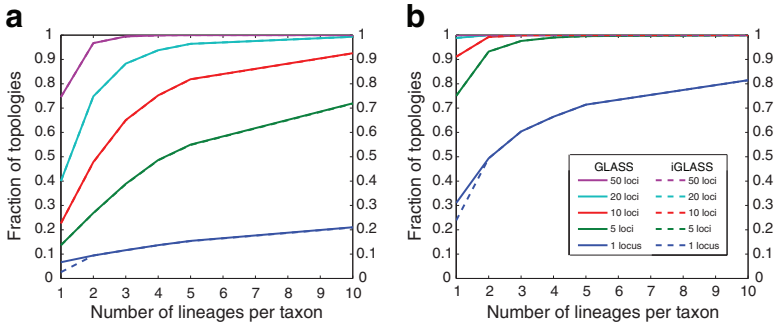

FIG. 7. The fraction of tree topologies correctly inferred by the approximate iGLASS and GLASS methods for two different five-taxon species trees. The tree in **(a)** is the same tree considered in Figure 6a,b. The tree in **(b)** is the same tree considered in Figure 6c,d. Plots show the fraction of 50,000 simulated data sets in which the species tree topology was correctly inferred by GLASS and approximate iGLASS.

D:0.5):0.025, **C**:0.525):0.025, **B**:0.55):10.0, **A**:10.55) and ((((**E**:0.5, **D**:0.5):0.2, **C**:0.7):1, **B**:1.7):10.0, **A**:11.7) that we considered, the GLASS and iGLASS methods have identical accuracies for estimating the topology. However, for the case in which only one lineage is sampled at only one locus, the GLASS method has slightly higher accuracy for inferring the topology (Fig. 7).

The reduction in accuracy for the case of one lineage and one locus was due to the fact that in this case, the iGLASS method estimated more than one pairwise divergence time in the species tree to be zero, resulting in ties that were sometimes resolved to produce a clade that was not on the true species tree. Multiple estimates of zero were produced in this case because the smallest possible expected value $E_0[\hat{t}_{AB}]$ of the GLASS estimate for a pair of taxa was equal to one, which was greater than at least two of the node heights in each tree that we considered (0.5 and 0.525 for the first tree, and 0.5 and 0.7 for the second tree).

For all other parameter values we considered, $E_0[\hat{t}_{AB}]$ was smaller than all of the node heights in either tree, and no estimates of zero were produced. For example, when two lineages were sampled per taxon, the smallest possible expected GLASS estimate was $E_0[T_{AB}] = 0.39$, which is smaller than 0.5, the smallest node height in either tree. Similarly, when one lineage was sampled per taxon at 5 loci, the smallest expected interspecific coalescence time was $E_0[T_{AB}] = 0.2$. Consequently, for all cases we considered except for the case of one sampled lineage per taxon at one locus, the accuracy of the iGLASS method for estimating topologies was the same as that of the GLASS method.

10. DISCUSSION

For two taxa, A and B, we have derived a closed-form expression for the distribution of $V_{AB} = \min_\ell \hat{t}_{AB}^{(\ell)} - \tau_{AB}$, the waiting time to the first interspecific coalescence across L loci, measuring from the divergence time τ_{AB}. By computing the expectation $E_{\tau_{AB}}[V_{AB}]$, we constructed a correction to the GLASS estimator $\hat{t}_{AB} = \min_\ell \hat{t}_{AB}^{(\ell)}$ of pairwise divergence times, which we call the iGLASS estimator.

Maruvka et al. (2011) have demonstrated that simple functions of time t in a population of constant size can provide useful deterministic continuous approximations of the number of lineages remaining at time t under the standard coalescent model. By approximating the number of lineages at time t by φ_x^t, the expected number of lineages remaining at time t when x lineages are sampled at time $t = 0$ and when x is not necessarily an integer, we derived an approximation $\tilde{\tau}_{AB}$ to the exact iGLASS estimator $\hat{\tau}_{AB}$ that is faster to compute than the exact value, and that is quite accurate even when the number of lineages is small.

Through simulations, we have shown that the exact and approximate iGLASS estimators reduce the bias in the GLASS estimates of pairwise divergence times. In addition, the exact iGLASS estimator $\hat{\tau}_{AB}$ and its approximation $\tilde{\tau}_{AB}$ generally reduce the mean squared error in the GLASS estimate of pairwise divergence times by approximately one half. This reduction accords with a theoretical prediction in the case in which a single lineage is sampled per taxon.

In our simulations, the accuracy of the iGLASS method for estimating topologies was similar to that of the GLASS method. In the case in which one lineage was sampled per taxon at one locus, iGLASS was slightly poorer, due to the fact that iGLASS produces divergence time estimates of zero whenever the GLASS estimate is smaller than its smallest possible expected value, $E_0[\hat{t}_{AB}]$. Because $E_0[\hat{t}_{AB}]$ is smaller when the number of sampled lineages or loci is larger, divergence time estimates of zero are less likely

when more lineages or loci are sampled. Therefore, the accuracy of the topology estimates produced by iGLASS are likely to be the same as those produced by GLASS whenever sufficiently many lineages or loci are sampled.

We have shown that the exact iGLASS estimator and its approximation are consistent estimators of the pairwise divergence time between a pair of taxa. Further, we have proven that applying any clustering method with nonzero L_∞-radius to the pairwise iGLASS estimates produces a statistically consistent estimator of the species tree topology.

Assuming that gene trees have been correctly inferred, the bias in the GLASS method itself decreases to zero quickly as the number of loci increases. Thus, our correction produces the greatest improvement when information is available for relatively few loci. As we have seen, however, the approximate iGLASS correction is fast to compute even for large numbers of loci, requiring only $O(n^2L|\mathcal{S}| + L|\mathcal{S}|^2)$ operations for a given level of precision, compared to $O(n^2L|\mathcal{S}|^2 + |\mathcal{S}|^3)$ operations for GLASS. Consequently, our new estimator provides a method that is reasonable to implement even when information is available at many loci.

11. APPENDIX A

A recursive formula for $E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B]$

In Appendix A, we derive Equation (5), the expected value of the difference $V_{AB} = \hat{t}_{AB} - \tau_{AB}$, conditional on the numbers of lineages that remain at each locus at the divergence time. Let C_ℓ be the event that the first coalescence occurs in locus ℓ ($\ell = 1, \ldots, L$). We then recursively consider what happens on the next coalescent event:

$$
\begin{aligned}
&E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B] \\
&\quad = \sum_{\ell=1}^{L} E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B, C_\ell] Pr(C_\ell) \\
&\quad = \sum_{\ell=1}^{L} \left[\frac{1}{\sum_{\lambda=1}^{L} \binom{k_{A_\lambda} + k_{B_\lambda}}{2}} + E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A - \mathbf{e}_\ell, \mathbf{k}_B] \frac{\binom{k_{A_\ell}}{2}}{\binom{k_{A_\ell} + k_{B_\ell}}{2}} \right. \\
&\qquad \left. + E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B - \mathbf{e}_\ell] \frac{\binom{k_{B_\ell}}{2}}{\binom{k_{A_\ell} + k_{B_\ell}}{2}} \right] Pr(C_\ell) \\
&\quad = \frac{1}{\sum_{\ell=1}^{L} \binom{k_{A_\ell} + k_{B_\ell}}{2}} + \sum_{\ell=1}^{L} \left\{ \left[E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A - \mathbf{e}_\ell, \mathbf{k}_B] \frac{\binom{k_{A_\ell}}{2}}{\binom{k_{A_\ell} + k_{B_\ell}}{2}} \right. \right. \\
&\qquad \left. \left. + E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B - \mathbf{e}_\ell] \frac{\binom{k_{B_\ell}}{2}}{\binom{k_{A_\ell} + k_{B_\ell}}{2}} \right] \frac{\binom{k_{A_\ell} + k_{B_\ell}}{2}}{\sum_{\lambda=1}^{L} \binom{k_{A_\lambda} + k_{B_\lambda}}{2}} \right\} \\
&\quad = \frac{1}{\sum_{\ell=1}^{L} \binom{k_{A_\ell} + k_{B_\ell}}{2}} \left[1 + \sum_{\ell=1}^{L} \left[E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A - \mathbf{e}_\ell, \mathbf{k}_B] \binom{k_{A_\ell}}{2} \right. \right. \\
&\qquad \left. \left. + E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B - \mathbf{e}_\ell] \binom{k_{B_\ell}}{2} \right] \right].
\end{aligned}
$$

Above, λ is a ''dummy'' summation variable. The second equality can be understood as follows. Because the time to the first coalescent event at locus ℓ is exponentially distributed with mean $1/\binom{k_{A_\ell} + k_{B_\ell}}{2}$ ($\ell = 1, \ldots, L$), the time to the first coalescence at some locus is distributed as the minimum of L such random variables. Therefore, the expected time to the first coalescent event is $1/\sum_{\lambda=1}^{L} \binom{k_{A_\lambda} + k_{B_\lambda}}{2}$ coalescent units. We must always wait this long on average before the first interspecific coalescent event. Given that the first coalescence occurs at locus ℓ, if the coalescence occurs among lineages from taxon A, an event that occurs with probability $\binom{k_{A_\ell}}{2} / \binom{k_{A_\ell} + k_{B_\ell}}{2}$, we must wait on average an additional

$E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A - \mathbf{e}_\ell, \mathbf{k}_B]$ time units. Similarly, with probability $\binom{k_{B_\ell}}{2} / \binom{k_{A_\ell}+k_{B_\ell}}{2}$, we must wait an additional $E_{\tau_{AB}}[V_{AB}|\mathbf{k}_A, \mathbf{k}_B - \mathbf{e}_\ell]$ time units on average. Finally, if the first coalescence at locus ℓ is interspecific, an event that has probability $\frac{k_{A_\ell}k_{B_\ell}}{\binom{k_{A_\ell}+k_{B_\ell}}{2}}$, no further waiting is necessary.

In the third equality, the term $1/\sum_{\lambda=1}^{L}\binom{k_{A_\lambda}+k_{B_\lambda}}{2}$ does not depend on ℓ and can be brought outside. Additionally, because the time to the first coalescence at locus ℓ is exponentially distributed with mean $1/\binom{k_{A_\ell}+k_{B_\ell}}{2}$, the first coalescence occurs at locus ℓ with probability $Pr(C_\ell) = \binom{k_{A_\ell}+k_{B_\ell}}{2} / \sum_{\ell=1}^{L}\binom{k_{A_\ell}+k_{B_\ell}}{2}$.

12. APPENDIX B

Derivation of Equation (7)

In Appendix B, we rely on results from Rosenberg (2003) to derive the probability distribution of M, the number of coalescent events up to and including the first interspecific coalescence, counting backwards in time from the divergence time.

Suppose that k_A and k_B lineages from taxa A and B, respectively, remain at time τ_{AB}. Equation (A8) of Rosenberg (2003) gives the probability $F_w^{A,B}(k_A, k_B)$ that an interspecific coalescence occurs among these lineages on or before the $(k - w)$th coalescence, where $k = k_A + k_B$. This probability is

$$F_w^{A,B}(k_A, k_B) = \sum_{x=\max(1,\, w+1-k_B)}^{k_A} \sum_{y=\max(1,\, w+1-x)}^{k_B} \frac{I_{k_A,x} I_{k_B,y}}{I_{k_A+k_B,w}} W(k_A - x, k_B - y)xyI_{x+y-1,w}, \tag{B.1}$$

where $I_{n,k} = [n!(n-1)!]/[2^{n-k}k!(k-1)!]$ (Rosenberg, 2003) is the number of ways in which n lineages can coalesce down to k lineages, and $W(i,j) = \binom{i+j}{i}$ (Rosenberg, 2003) is the number of ways of ''interweaving'' the coalescent events among lineages only from taxon A with the coalescent events among lineages only from taxon B.

Each term in the summation (B.1) is the joint probability that the first interspecific coalescence occurs when the k_A and k_B lineages have x and y ancestors, respectively, and that the first interspecific coalescence occurs on or before the $(k - w)$th coalescence. If $w = 1$, then each term is just the probability that the first interspecific coalescence occurs when the k_A and k_B lineages have x and y ancestors.

Since different choices of x and y (say, (x_1, y_1) and (x_2, y_2) where $x_1 \neq x_2$ or $y_1 \neq y_2$, or both) correspond to mutually exclusive events, and since the sum $x + y$ specifies M through the relationship $x + y = k - M + 1$, to derive the probability that the first interspecific coalescence is the Mth coalescence (Equation 7), we can set $w = 1$ and sum over all x and y such that $x + y = k - M + 1$, i.e., over all mutually exclusive events corresponding to the case in which the first interspecific coalescence is the Mth coalescence.

To determine the values of x and y corresponding to the case $M = m$, we can write $x = k - m + 1 - y$. Note that x is at most k_A and at least 1, and thus, $1 \leq x \leq \min\{k - m, k_A\}$. Similarly, by symmetry in x and y, $1 \leq y \leq \min\{k - m, k_B\}$, giving $x = k - m + 1 - y \geq k - m + 1 - \min\{k - m, k_B\} = \max\{1, k_A - m + 1\}$. This inequality yields the constraint $\max\{1, k_A - m + 1\} \leq x \leq \min\{k - m, k_A\}$. Thus, we obtain

$$\begin{aligned} &Pr(M = m) \\ &\quad = \sum_{x=\max\{1,\, k_A-m+1\}}^{\min\{k-m,\, k_A\}} \frac{I_{k_A,x} I_{k_B,k-m+1-x}}{I_{k_A+k_B,1}} W(k_A - x, k_B - y)x(k - m + 1 - x)I_{k_A+k_B-m,1.} \end{aligned}$$

Making the change of variables $\eta = k_A - x$ and noting that $k_B - y = m - 1 - \eta$ because $k_A + k_B - m + 1 = x + y$, we get

$$\begin{aligned} &Pr(M = m) \\ &\quad = \frac{I_{k_A+k_B-m,1}}{I_{k_A+k_B,1}} \sum_{\eta=\max\{0,\, m-k_B\}}^{\min\{m-1,\, k_A-1\}} W(\eta, m - 1 - \eta)I_{k_{A,}\ k_A-\eta} I_{k_{B,}\ k_B-m+1+\eta} \\ &\qquad \times (k_A - \eta)(k_B - m + 1 + \eta). \end{aligned}$$

Using $I_{n,k} = [n!(n-1)!]/[2^{n-k}k!(k-1)!]$ and $W(i,j) = \binom{i+j}{i}$, we get

$$
\begin{aligned}
&Pr(M=m)\\
&\quad=\frac{I_{k_A+k_B-m,1}}{I_{k_A+k_B,1}}\sum_{\eta=\max\{0,\,m-k_B\}}^{\min\{m-1,\,k_A-1\}}\binom{m-1}{\eta}\frac{k_A!(k_A-1)!}{2^{k_A-(k_A-\eta)}(k_A-\eta)!(k_A-\eta-1)!}\\
&\quad\times\frac{k_B!(k_B-1)!}{2^{k_B-(k_B-m+1+\eta)}(k_B-m+1+\eta)!(k_B-m+\eta)!}\\
&\quad\times(k_A-\eta)(k_B-m+1+\eta)\\
&\quad=\frac{I_{k_A+k_B-m,1}}{I_{k_A+k_B,1}}\sum_{\eta=\max\{0,\,m-k_B\}}^{\min\{m-1,\,k_A-1\}}\binom{m-1}{\eta}\frac{k_A!^2}{k_A(k_A-\eta-1)!^2}\frac{k_B!^2}{k_B(k_B-m+\eta)!^2}\frac{1}{2^{m-1}}\\
&\quad=\frac{I_{k_A+k_B-m,1}}{2^{m-1}k_Ak_BI_{k_A+k_B,1}}\sum_{\eta=\max\{0,\,m-k_B\}}^{\min\{m-1,\,k_A-1\}}\binom{m-1}{\eta}k^2_{A_{[\eta+1]}}k^2_{B_{[m-\eta]}},
\end{aligned}
\tag{B.2}
$$

where $k_{[i]}=k!/(k-i)!$.

When either $k_A=1$ or $k_B=1$, Equation (B.2) has a particularly simple form. Without loss of generality, suppose that $k_B=1$. Then $\max\{0,\ m-k_B\}=m-1$ because $m\geq 1$, and $\min\{m-1,\ k_A-1\}=m-1$ because $m\leq k_A+k_B-1=k_A$. Therefore, using $k=k_A+1$, Equation (B.2) simplifies as follows:

$$
\begin{aligned}
Pr(M=m)&=\frac{1}{k_A2^{m-1}}\frac{2^{k_A+1-1}}{(k_A+1)!k_A!}\frac{(k_A+1-m)!(k_A+1-m-1)!}{2^{k_A+1-m-1}}k^2_{A[m]}1^2_{[1]}\\
&=\frac{2(k_A+1-m)}{k_A(k_A+1)}\frac{(k_A-m)!(k_A-m)!}{k_A!k_A!}\frac{k_A!^2}{(k_A-m)!^2}\\
&=\frac{2(k-m)}{k(k-1)}.
\end{aligned}
\tag{B.3}
$$

13. APPENDIX C.

Computational complexity of approximate iGLASS

We now compute the computational complexity of the approximate iGLASS method, Equation (21). To compute the iGLASS correction for each pair of taxa X and Y in $\mathcal{S}$, we first evaluate Equation (20) for many different values of τ_{XY}. In particular, to numerically obtain the inverse in Equation (21), we compute Equation (20) for each divergence time estimate τ_{XY} in the set $\{i\Delta t\}_{i=0}^{\Gamma_{XY}}$, where Δt is a fixed time-step and $\Gamma_{XY}=\lceil\hat{t}_{XY}/\Delta t\rceil$. We then estimate τ_{XY} by the value $\tau\in\{i\Delta t\}_{i=0}^{\Gamma_{XY}}$ that minimizes the quantity $|\tau+\tilde{E}_\tau[V_{XY}]-\hat{t}_{XY}|$.

To evaluate the integral in Equation (20), we assume that numerical integration is carried out by computing the Riemann sum with fixed step-size Δt. We truncate the outer integral at $P\Delta t$, where P is large enough that the tail of the outer integral in Equation (20) is smaller than some predefined value $\epsilon>0$. For a given value of ϵ, a sufficiently-large value of P can be found by bounding the integral in Equation (20). The bound can be obtained by noting that $\varphi^z_n\geq 1$ for all n and z in $\mathbb{R}^+$, and thus, the integrand in Equation (20) is smaller than $\exp\{-Lt\}$, which is easily integrated. Converting the integrals in Equation (20) to summations gives

$$
\begin{aligned}
\tilde{E}_{\tau_{XY}}[V_{XY}]&\approx\sum_{\alpha=0}^{P}\Delta t\exp\left\{-\sum_{\ell=1}^{L}\int_{z=0}^{\alpha\Delta t}\varphi^z_{E_{\tau_{XY}}[K_{X_\ell}]}\varphi^z_{E_{\tau_{XY}}[K_{Y_\ell}]}dz\right\}\\
&\approx\sum_{\alpha=0}^{P}\Delta t\exp\left\{-\sum_{\ell=1}^{L}\sum_{\beta=0}^{\alpha}\Delta t\ \varphi^{\beta\Delta t}_{E_{\tau_{XY}}[K_{X_\ell}]}\varphi^{\beta\Delta t}_{E_{\tau_{XY}}[K_{Y_\ell}]}\right\}.
\end{aligned}
\tag{C.1}
$$

Once $\varphi^{\beta\Delta t}_{E_{\tau_{XY}}[K_{X_\ell}]}$ and $\varphi^{\beta\Delta t}_{E_{\tau_{XY}}[K_{Y_\ell}]}$ have been pre-computed and stored for all values at which they are evaluated in the summation, the exponent in Equation (C.1) requires $O(L\alpha)$ operations, where α is the index in the outermost summation. Thus, we have the following result:

After pre-computing the terms in the summand, the summation (C.1) requires $O(LP^2)$ operations.

For each taxon $X \in \mathcal{S}$, let $\Gamma_X = \max_{Y \in \mathcal{S}, Y \neq X} \Gamma_{XY}$; in other words, $\Gamma_X \Delta t$ is, to precision Δt, the maximum pairwise divergence time between taxon X and any other taxon. For each $\ell = 1, \ldots, L$ and for each $X \in \mathcal{S}$, we must ultimately compute $E_\tau[K_{X_\ell}]$ for each $\tau \in \{i\Delta t\}_{i=0}^{\Gamma_X}$, and we must compute $\varphi^t_{E_\tau[K_{X_\ell}]}$ for each $t \in \{i\Delta t\}_{i=0}^{P}$ and for each $\tau \in \{j\Delta t\}_{j=0}^{\Gamma_X}$. However, note that $\varphi_n^{t+\Delta t} = \varphi^{\Delta t}_{\varphi_n^t}$, and note that $\varphi^\tau_{n_{X_\ell}} = E_\tau[K_{X_\ell}]$ by definition (Equations (15) and (16)). Therefore, we have $\varphi^{i\Delta t}_{E_\tau[K_{X_\ell}]} = \varphi^{i\Delta t}_{\varphi^\tau_{n_{X_\ell}}} = \varphi^{\tau + i\Delta t}_{n_{X_\ell}}$ for all $i \in \mathrm{Z}^+$, and thus, it suffices to pre-compute $\varphi^\tau_{n_{X_\ell}}$ for all $\tau \in \{j\Delta t\}_{j=0}^{\Gamma_X + P}$ for each $X \in \mathcal{S}$ and for each ℓ ($\ell = 1, \ldots, L$).

Let $n = \max_{X \in \mathcal{S}} \max_{\ell \in \{1, \ldots, L\}} n_{X_\ell}$ be the maximal number of lineages sampled from any taxon, and let $Q = P + \max_{X \in \mathcal{S}} \Gamma_X$. Then for a given taxon $X \in \mathcal{S}$ and for a given ℓ ($\ell = 1, \ldots, L$), the amount of time needed to compute $\varphi^\tau_{n_{X_\ell}}$ for all $\tau \in \{j\Delta t\}_{j=0}^{\Gamma_X + P}$ is bounded by the time needed to compute φ_n^t for all $t \in \{i\Delta t\}_{i=0}^{Q}$.

Because the summand in Equation (15) requires $O(n)$ operations, (a rising factorial and a falling factorial totaling n multiplications), computing Equation (15) for a given value of t requires $O(n^2)$ operations. Therefore, evaluating (15) for each time t in $\{i\Delta t\}_{i=0}^{Q}$ requires $O(n^2 Q)$ operations, and pre-computing φ_n^t for all $t \in \{i\Delta t\}_{i=0}^{Q}$ also requires $O(n^2 Q)$ operations. This gives the following result:

> Pre-computing $\varphi^\tau_{n_{X_\ell}}$ for all $\tau \in \{j\Delta t\}_{j=0}^{\Gamma_X + P}$ for all $X \in \mathcal{S}$ and for all ℓ ($\ell = 1, \ldots, L$) requires $O(n^2 QL|\mathcal{S}|)$ operations.

Once all values of $\varphi^\tau_{n_{X_\ell}}$ have been pre-computed and stored, Equation (C.1) must be computed for each $\tau \in \{0, 1, \ldots, \Gamma_{XY}\}$ for each pair of taxa $X, Y \in \mathcal{S}$. Equation (C.1) requires $O(LP^2)$ operations for each value of τ. Therefore, because $\Gamma_{XY} \leq Q$, computing (C.1) for a pair of taxa requires $O(LP^2 Q)$ operations. Because $P \leq Q$, this simplifies to $O(LQ^3)$ operations. Therefore, computing (C.1) for all $\binom{|\mathcal{S}|}{2}$ pairs of taxa requires $O(LQ^3|\mathcal{S}|^2)$ operations. Combining this quantity with the number of operations necessary to pre-compute the values of $\varphi^\tau_{n_{X_\ell}}$ gives the following result:

> Including all pre-computations, the total number of operations required to compute Equation (C.1) for all $\binom{|\mathcal{S}|}{2}$ pairs of taxa is $O(n^2 QL|\mathcal{S}| + LQ^3|\mathcal{S}|^2)$.

Note that once all values of $\varphi^\tau_{n_{X_\ell}}$ have been pre-computed and stored, the cost of computing (C.1) does not depend on the magnitude of the n_{X_ℓ}, only on the number of terms in the summation. Thus, the complexity only depends on n through the pre-computation step.

The only other computations needed to compute the approximate iGLASS correction are those associated with finding $\arg\min_\tau |\tau + \tilde{E}_\tau[V_{AB}] - \hat{t}_{AB}|$ and those associated with the single-linkage clustering step. We must perform $\binom{|\mathcal{S}|}{2}$ searches to find the value of τ that minimizes $|\tau + \tilde{E}_\tau[V_{AB}] - \hat{t}_{AB}|$ for each of the $\binom{|\mathcal{S}|}{2}$ pairs of taxa. An exhaustive search is bounded by the number of values of τ, which is always less than or equal to Q. Thus, correcting the GLASS method requires $O(Q|\mathcal{S}|^2)$ operations. Finally, single-linkage clustering requires at most $O(|\mathcal{S}|^2)$ operations (Gordon, 1996). Thus, the entire correction procedure requires $O(n^2 QL|\mathcal{S}| + LQ^3|\mathcal{S}|^2 + Q|\mathcal{S}|^2 + |\mathcal{S}|^2)$ operations. Terms can be combined to get the following result:

> The entire approximate iGLASS correction procedure requires $O(n^2 QL|\mathcal{S}| + LQ^3|\mathcal{S}|^2)$ operations.

It is useful to compare the complexity of approximate iGLASS to the complexity of GLASS for a given precision. The choices of Δt and P determine the precision in computing the approximate iGLASS correction, in other words, the error between the outcome of the numerical steps that we have just outlined, and the outcome of exactly computing Equation (20) and exactly solving Equation (21). Together, Δt and P determine $Q = P + \lceil \max_{XY \in \mathcal{S}} \hat{t}_{XY} / \Delta t \rceil$. Thus, Q is a tuning parameter that affects the precision in our numerical steps. For fixed Q, the complexity of approximate iGLASS is $O(n^2 L|\mathcal{S}| + L|\mathcal{S}|^2)$. In comparison, a similar analysis demonstrates that the GLASS method requires $O(n^2 L|\mathcal{S}|^2 + |\mathcal{S}|^3)$ operations.

14. APPENDIX D

Consistency of GLASS for divergence times

Mossel and Roch (2010) proved that the GLASS method is a consistent estimator of the species tree topology as the number of loci approaches infinity. Liu et al. (2010) proved that the GLASS estimator is consistent for pairwise divergence times in the case in which a single lineage is sampled per taxon.

Here, we prove that GLASS is a consistent estimator of pairwise divergence times in the case in which arbitrarily many lineages are sampled per taxon. Our argument is a minor extension of the consistency proof in Liu et al. (2010).

Theorem D.1. *Consider two taxa, A and B, with divergence time τ_{AB}. The GLASS estimator $\hat{t}_{AB}$ is a consistent estimator of τ_{AB}.*

Proof. At each locus ℓ ($\ell = 1, \ldots, L$), consider a lineage a_ℓ sampled at random from taxon A and a lineage b_ℓ sampled at random from taxon B. The time V_{AB}^{ℓ} to the first interspecific coalescence at locus ℓ is less than or equal to the coalescence time between a_ℓ and b_ℓ, which we denote by V_{a_ℓ, b_ℓ}. Therefore, using the fact that the GLASS estimate is given by $\hat{t}_{AB} = \tau_{AB} + V_{AB}$, and following Liu et al. (2010), we obtain $Pr(|\hat{t}_{AB} - \tau_{AB}| > \epsilon) = Pr(V_{AB} > \epsilon) = Pr(\min_\ell V_{AB}^{\ell} > \epsilon) \leq Pr(\min_\ell V_{a^\ell, b^\ell} > \epsilon) = \prod_{\ell=1}^{L} Pr(V_{a^\ell, b^\ell} > \epsilon) = e^{-L\epsilon}$. Here, to obtain the last equality, we have used the fact that V_{a^ℓ, b^ℓ} is exponentially distributed with mean 1 coalescent unit of N generations. Thus, we have

$$0 \leq Pr(|\hat{t}_{AB} - \tau_{AB}| > \epsilon) \leq e^{-L\epsilon},$$

from which it follows that $Pr(|\hat{t}_{AB} - \tau_{AB}| > \epsilon) \to 0$ as $L \to \infty$ by the ''squeeze theorem.'' ■

15. APPENDIX E

iGLASS and approximate iGLASS are consistent estimators of pairwise divergence times

Here, we prove that the expectation $E_{\tau_{AB}}[V_{AB}]$ of the difference $V_{AB} = \hat{t}_{AB} - \tau_{AB}$ between the GLASS estimate $\hat{t}_{AB}$ and the divergence time τ_{AB} is bounded above by $1/L$. Thus, $E_{\tau_{AB}}[V_{AB}] \to 0$ as $L \to \infty$. Using Equation (1), we then show that the difference between the GLASS estimator and the iGLASS estimator is bounded above by $1/L$. Thus, the difference goes to 0 as $L \to \infty$. A similar result is proven for the expectation $\tilde{E}_{\tau_{AB}}[V_{AB}]$ used in the approximate iGLASS correction (Equation 21).

Since GLASS is a consistent estimator of pairwise divergence times, these results can be used to show that exact iGLASS and approximate iGLASS are consistent estimators of pairwise divergence times, as they converge to the same limit as the GLASS estimator in the limit $L \to \infty$.

Lemma E.1. *For taxa A and B, let $E_{\tau_{AB}}[V_{AB}]$ be the expectation of the difference $V_{AB} = \hat{t}_{AB} - \tau_{AB}$ between the GLASS estimate $\hat{t}_{AB}$ and the divergence time τ_{AB}. Then $E_{\tau_{AB}}[V_{AB}] \leq 1/L$.*

Proof. In Theorem D.1, we saw that $0 \leq Pr(|\hat{t}_{AB} - \tau_{AB}| > \epsilon) \leq e^{-L\epsilon}$ for all $\epsilon > 0$. Thus,

$$\begin{aligned} E_{\tau_{AB}}[V_{AB}] &= \int_{v=0}^{\infty} Pr(V_{AB} > v) dv \\ &= \int_{v=0}^{\infty} Pr(|\hat{t}_{AB} - \tau_{AB}| > v) dv \\ &\leq \int_{v=0}^{\infty} e^{-Lv} dv \\ &= \frac{1}{L}, \end{aligned}$$

proving the result. ■

Lemma E.2. *The approximation $\tilde{E}_{\tau_{AB}}[V_{AB}]$ satisfies $\tilde{E}_{\tau_{AB}}[V_{AB}] \leq 1/L$.*

Proof. For any n and t, the expected number of lineages φ_n^t remaining at any given time t is at least 1. Therefore, for any ℓ ($\ell = 1, \ldots, L$)

$$\int_{z=0}^{t} \varphi^z_{E_{\tau_{AB}}[K_{A_\ell}]} \varphi^z_{E_{\tau_{AB}}[K_{B_\ell}]} dz \geq t.$$

Consequently,

$$\begin{aligned}
\tilde{E}_{\tau_{AB}}[V_{AB}] &= \int_{t=0}^{\infty} \exp\left\{ -\sum_{\ell=1}^{L} \int_{z=0}^{t} \varphi^z_{E_{\tau_{AB}}[K_{A_\ell}]} \varphi^z_{E_{\tau_{AB}}[K_{B_\ell}]} dz \right\} dt \\
&\leq \int_{t=0}^{\infty} \exp\left\{ -\sum_{\ell=1}^{L} t \right\} dt \\
&= \int_{t=0}^{\infty} e^{-Lt} dt \\
&= \frac{1}{L},
\end{aligned}$$

proving the result. ■

The following corollary proves that after the correction procedure (Equation 2), both the exact and approximate iGLASS estimates differ from the GLASS estimate by at most $1/L$ coalescent units.

Corollary E.3. *For two taxa A and B, let $C_{AB} = \hat{t}_{AB} - \hat{\tau}_{AB}$ and $\tilde{C}_{AB} = \hat{t}_{AB} - \tilde{\tau}_{AB}$ be the differences between the GLASS estimate and the exact and approximate iGLASS estimates, respectively. Then $C_{AB} \leq 1/L$ and $\tilde{C}_{AB} \leq 1/L$.*

Proof. Using Equation (2), if $E_0[V_{AB}] \leq \hat{t}_{AB}$, then the iGLASS estimate is obtained by solving $\tau_{AB} = \hat{t}_{AB} - E_{\tau_{AB}}[V_{AB}]$ for τ_{AB}. In this case, the difference C_{AB} is at most $1/L$ by Lemma E.1. On the other hand, if $\hat{t}_{AB} \in [0, E_0[V_{AB}])$, then the iGLASS estimate is given by $\hat{\tau}_{AB} = 0$. Since $[0, E_0[V_{AB}]) \subseteq [0, 1/L)$ by Lemma E.1, we have $|\hat{t}_{AB} - \hat{\tau}_{AB}| < 1/L$. Thus, in both cases, $C_{AB} \leq 1/L$.

The same argument using Lemma E.2 and Equation (21) rather than Lemma E.1 and Equation (2) establishes $\tilde{C}_{AB} \leq 1/L$, proving the result. ■

ACKNOWLEDGMENTS

We are grateful to Michael DeGiorgio, Lucy Huang, and Laura Helmkamp for helpful discussions, and to Lucy Huang for suggesting the name iGLASS. We also thank two anonymous reviewers for their careful reading and helpful suggestions, and for simplifying the proofs of Theorem D.1 and Lemma E.1. This work was supported by the NSF (grants DEB-0716904 and DBI-1146722), by a grant from the Burroughs Wellcome Fund, and by the NIH (training grant T32 HG00040).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Atteson, K. 1999. The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica* 25, 251–278.

Casella, G., and Berger, R.L. 2002. *Statistical Inference,* 2nd ed. Duxbury Press, Pacific Grove, CA.

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

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

Edwards, S.V., and Beerli, P. 2000. Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* 54, 1839–1854.

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

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

Gascuel, O., and McKenzie, A. 2004. Performance analysis of hierarchical clustering algorithms. *J. Classif.* 21, 3–18.

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

Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model. *Bioinformatics* 18, 337–338.

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.

Liu, L., Yu, L., Kubatko, L., et al. 2009a. Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.* 53, 320–328.

Liu, L., Yu, L., Pearl, D.K., et al. 2009b. Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* 58, 468–477.

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

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

Maruvka, Y.E., Shnerb, N.M., Bar-Yam, Y., et al. 2011. Recovering population parameters from a single gene genealogy: an unbiased estimator of the growth rate. *Mol. Biol. Evol.* 28, 1617–1631.

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

Nichols, R. 2001. Gene trees and species trees are not the same. *Trends Ecol. Evol.* 16, 358–364.

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

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

Rosenberg, N.A. 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution* 57, 1465–1477.

Rosenberg, N.A., and Feldman, M.W. 2002. The relationship between coalescence times and population divergence times, 130–164. *In* Slatkin, M., Veuille, M., eds. *Modern Developments in Theoretical Population Genetics*. Oxford University Press, Oxford, UK.

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

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Semple, C., and Steel, M. 2003. *Phylogenetics*. Oxford University Press, New York.

Sneath, P.H.A. 1957. The application of computers to taxonomy. *J. Gen. Microbiol.* 17, 201–226.

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

Sørensen, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *K. Dan. Vidensk. Selskab Biol. Skrift.* 5, 1–34.

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

Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* 26, 119–164.

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

Address correspondence to:
Ethan M. Jewett
Department of Biology
Stanford University
371 Serra Mall
Stanford, CA 94305-5020

E-mail: emjewett@stanford.edu