Maximum Likelihood Estimation of Species Trees from Gene Trees in the Presence of Ancestral Population Structure

Hillary Koch[1] and Michael DeGiorgio[2,*]

[1]Department of Statistics, Pennsylvania State University

[2]Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University

*Corresponding author: E-mail: mdegiorg@fau.edu.

Accepted: January 23, 2020

Abstract

Though large multilocus genomic data sets have led to overall improvements in phylogenetic inference, they have posed the new challenge of addressing conflicting signals across the genome. In particular, ancestral population structure, which has been uncovered in a number of diverse species, can skew gene tree frequencies, thereby hindering the performance of species tree estimators. Here we develop a novel maximum likelihood method, termed TASTI (Taxa with Ancestral structure Species Tree Inference), that can infer phylogenies under such scenarios, and find that it has increasing accuracy with increasing numbers of input gene trees, contrasting with the relatively poor performances of methods not tailored for ancestral structure. Moreover, we propose a supertree approach that allows TASTI to scale computationally with increasing numbers of input taxa. We use genetic simulations to assess TASTI's performance in the three- and four-taxon settings and demonstrate the application of TASTI on a six-species Afrotropical mosquito data set. Finally, we have implemented TASTI in an open-source software package for ease of use by the scientific community.

Key words: consistency, discordance, phylogenetics, structure, TASTI.

Introduction

Large multilocus data sets are becoming ever more common in systematic biology (Cranston et al. 2009; Song et al. 2012; Song et al. 2012; Salichos and Rokas 2013; DeGiorgio et al. 2014; Fontaine et al. 2015; Peters et al. 2017). However, even with these more substantial data sets, the presence of gene tree discordance, in which inferred gene tree topologies conflict across loci, can be problematic when making phylogenetic inference. A major source of gene tree discordance is incomplete lineage sorting, which occurs when sampled lineages fail to coalesce, or find a common ancestor, in the first population in which they are able to do so. Accordingly, several statistically consistent methods for species tree inference that are robust to incomplete lineage sorting under the multispecies coalescent model have been developed (Kubatko et al. 2009; Degnan et al. 2009; Liu, Yu, and Pearl 2010; Mossel and Roch 2010; Helmkamp et al. 2012; Jewett and Rosenberg 2012; Wu 2012; Mirarab et al. 2014; Mirarab and Warnow 2015).

The multispecies coalescent assumes that each modern and ancestral species is unstructured and has a constant population size and that each pair of lineages within a given ancestral species has an equal probability of coalescing (Nakhleh 2013). Under these assumptions, incomplete lineage sorting leads to symmetries in gene tree distributions for any species tree, regardless of the number of taxa. For example, if a pair of taxa A and B are sister species on a species tree, then for all other species *X*, species A and *X* and species B and *X* have the same probability of being sister taxa in a gene tree (Allman et al. 2011). In the face of some form of gene flow between species, such as hybridization, continuous migration, or horizontal gene transfer, asymmetries in gene tree distributions can arise (Yu et al. 2011, 2012; Leaché et al. 2014; Solís-Lemus et al. 2016; Tian and Kubatko 2016; Long and Kubatko 2018). As such, asymmetries in gene tree distributions are often attributed to gene flow between species (McGuire et al. 2007; Escobar et al. 2012; Marcussen et al. 2014).

Despite the common crediting of asymmetries in gene tree distribution to interspecies gene flow, they can also emerge in the absence of such gene flow when ancestral populations are structured (Slatkin and Pollack 2008). In this case, gene flow does not occur between taxa, but rather between demes within taxa. There are several hypothesized examples of structured ancestral species (Garrigan et al. 2005; Thalmann et al. 2006; White et al. 2009), and genomic signatures of ancestral structure have been uncovered in a number of diverse lineages, including mouse (White et al. 2009) and yeast (Yu et al. 2012). Still, many methods for inferring species trees from multilocus data are not robust to ancestral structure and can be proven to be positively misleading (DeGiorgio and Rosenberg 2016). DeGiorgio and Rosenberg (2016) demonstrate that exceptions to this rule are the maximum likelihood estimators of the species tree implemented in GLASS (Mossel and Roch 2010), STEM (Kubatko et al. 2009), and Maximum Tree (Liu, Yu, and Pearl 2010); however, these algorithms underperform on empirical data when gene trees are inferred rather than known.

Here we detail an algorithm named TASTI (Taxa with Ancestral structure Species Tree Inference), a maximum likelihood method for inferring three-taxon species tree topologies when ancestral species are structured. We demonstrate that by explicitly incorporating the potential for population structure into a multispecies coalescent framework, the method is robust to population structure, and its performance typically does not significantly erode when applied to empirical data. Finally, we propose an approach to infer species tree topologies under ancestral population structure for an arbitrarily large number of species. We acknowledge that our method, which generalizes the standard multispecies coalescent, contains its own oversimplifications of ancestry, but maintain that it is a step in relaxing the common assumption of unstructured ancestral populations used in phylogenetics. We show via simulations in the three- and four-taxon settings that TASTI outperforms competing methods MP-EST (Liu, Yu, and Edwards 2010), STELLS2 (Pei and Wu 2017), and STEM2.0 (Kubatko et al. 2009, hereafter referred to as STEM) when population structure is present, but remains competitive under the unstructured multispecies coalescent.

Materials and Methods

Modeling Three Species with Ancestral Population Structure

In this section, we describe a model relating three species with population structure in each ancestral population and use it to construct probabilities of gene trees given a set of model parameters. In our model, the topologies $((ab)c)$ and ((AB)C) denote a gene tree and species tree, respectively, with sister species A and B and outgroup C. Figure 1*a* depicts a species tree with topology ((AB)C), in a configuration such that species B and C descend from the same ancestral subpopulation.

The species tree in our two-subpopulation model is characterized as follows. Divergence times, measured in expected number of mutations per site, are denoted $\boldsymbol{\tau} = (\tau_1, \tau_2)$, where the internal branch length $\tau_2 - \tau_1$ is of particular interest. For all ancestral species X, $\theta_{X,j} = 4N_{X,j}\mu$, where $N_{X,j}$ is the effective population size of subpopulation j in ancestral species X, and μ is the mutation rate per site per generation. The mutation-scaled coalescent rates within subpopulations of ancestral species are inversely proportional to their respective values described in the vector $\boldsymbol{\theta} = (\theta_{\text{AB},1}, \theta_{\text{AB},2}, \theta_{\text{ABC},1}, \theta_{\text{ABC},2})$. Migration occurs between subpopulations at mutation-scaled rates according to the vector $\mathbf{M} = (M_{\text{AB},1}, M_{\text{AB},2}, M_{\text{ABC},1}, M_{\text{ABC},2})$, where $M_{X,j}$ is the rate at which lineages move to subpopulation j when in ancestral species X. Specifically, we define $M_{X,j} = 4N_{X,j}m_{X,j}/\theta_{X,j}$, where $m_{X,j}$ is the fraction of subpopulation j in ancestral species X made up of migrants from its complementary subpopulation each generation.

Our method partitions the species tree into three distinct time transects on the intervals $[0, \tau_1)$, $[\tau_1, \tau_2)$, and $[\tau_2, \infty)$. During the interval $[0, \tau_1)$, we assume extant populations to be unstructured. Along the internal branch $[\tau_1, \tau_2)$, we introduce ancestral subpopulations such that migration and coalescence events may occur between the two lineages from sister species A (originating in subpopulation 1) and B (originating in subpopulation 2). In the final interval $[\tau_2, \infty)$, the lineage from species C is introduced into subpopulation 2. This setting matches the one presented by Slatkin and Pollack (2008), up to the first coalescence event. Migration and coalescence events occur until lineages from each species find their most recent common ancestor (MRCA), which may only happen between two lineages when they are in the same subpopulation. Since we are in essence estimating the waiting times until some occurrence, the distribution of migration and coalescent events can be modeled using an exponential distribution parametrized by an instantaneous rate matrix (Hobolth et al. 2011; Tian and Kubatko 2016). This parameterization yields the matrix exponential $e^{\mathbf{Q}} = \sum_{k=0}^{\infty} \mathbf{Q}^k/k!$, whose (i, j)th entry is denoted $(e^{\mathbf{Q}})_{ij}$.

Instantaneous Rate Matrices for the Model

The migration and coalescence of lineages within a structured population can be explicitly modeled using instantaneous rate matrices. For time to the first coalescence $T_1 \in [\tau_1, \tau_2)$, where migration and coalescence can only occur for a pair of lineages, one from species A and one from species B, we define the instantaneous rate matrix $\mathbf{Q}_{\text{AB}}$.

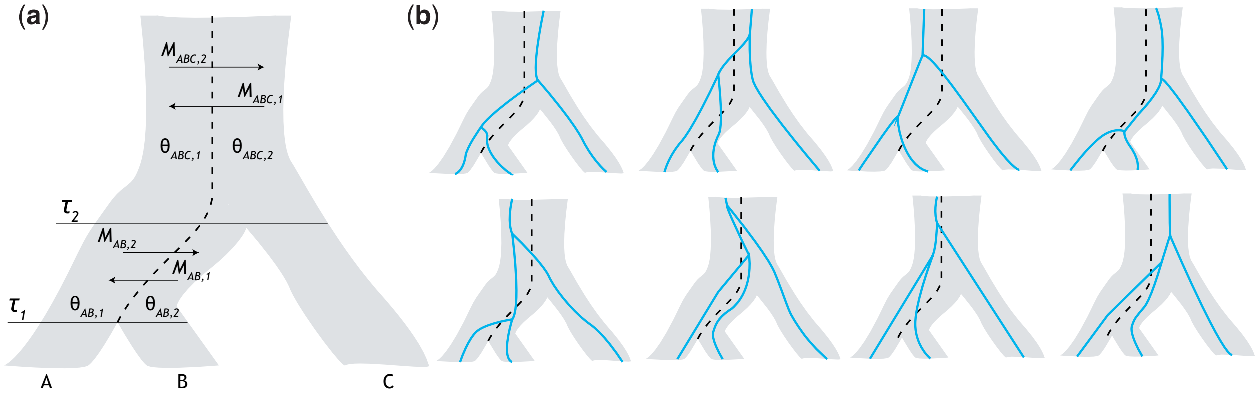

FIG. 1.—Modeling ancestral population structure. (*a*) Model of the relationships among species A, B, and C with divergence times τ_1 and τ_2. Ancestral species belong to one of two subpopulations of scaled coalescent rates $\theta_{X,k}$ with migration between subpopulations at rates $M_{AB,1}$ and $M_{AB,2}$ below the root and $M_{ABC,1}$ and $M_{ABC,2}$ above the root. Lineages from species A descend from subpopulation 1, whereas lineages from species B and C descend from subpopulation 2. (*b*) Incorporating population structure into the model increases the number of possible paths lineages from species A, B, and C might take to find their most recent common ancestor. For example, assuming that lineages from species A merge into subpopulation 1 and lineages from species B and C merge into subpopulation 2, a gene tree where lineages from species A and B coalesce first may arise from eight possible histories, as depicted here.

$$\mathbf{Q}_{AB} := \begin{array}{c} \\ a_1b_2 \\ a_2b_1 \\ a_1b_1 \\ a_2b_2 \\ (ab)_1 \\ (ab)_2 \end{array} \begin{array}{c} \begin{array}{cccccc} a_1b_2 & a_2b_1 & a_1b_1 & a_2b_2 & (ab)_1 & (ab)_2 \end{array} \\ \left[\begin{array}{cccccc} - & 0 & M_{AB,1} & M_{AB,2} & 0 & 0 \\ 0 & - & M_{AB,1} & M_{AB,2} & 0 & 0 \\ M_{AB,2} & M_{AB,2} & - & 0 & c_{AB,1} & 0 \\ M_{AB,1} & M_{AB,1} & 0 & - & 0 & c_{AB,2} \\ 0 & 0 & 0 & 0 & - & M_{AB,2} \\ 0 & 0 & 0 & 0 & M_{AB,1} & - \end{array}\right] \end{array}. \tag{1}$$

In our notation, a_ib_j indicates that the lineage (denoted a) from species A is currently in subpopulation i and the lineage (denoted b) from species B is currently in subpopulation j, where $i,j \in \{1,2\}$. Further, $(ab)_i$ indicates that lineages a and b have coalesced and are in subpopulation i, where $i \in \{1,2\}$. We refer to these states in the coalescent history as states 1 through 6 (i.e., states a_1b_2 through $(ab)_2$) of $\mathbf{Q}_{AB}$. The coalescence rate in subpopulation i is $c_{AB,i} = 2/\theta_{AB,i}$, and the dashes along the diagonal represent the negative sum of the elements of the corresponding row, such that each row sums to zero (Kingman 1982). One can see that this matrix embeds several assumptions. First, we only allow for one event at any given instant. For example, we do not permit a lineage in one subpopulation and a lineage in another subpopulation to migrate simultaneously. Additionally, once lineages coalesce, they cannot uncoalesce. Under our model, at time τ_1, lineages from species A are in subpopulation 1, and lineages from species B are in subpopulation 2.

The interval $[\tau_2, \infty)$ is described using one of two possible instantaneous rate matrices. If the first coalescence is yet to occur, then the 20 × 20 matrix $\mathbf{Q}_{ABC}$ governs the dynamics. In the interest of space, this matrix is left to supplementary figure S1, Supplementary Material online. We should note, however, that because any two lineages are equally likely to coalesce in states when all three lineages are in subpopulation i, the rate of coalescence in subpopulation i in these scenarios is $\binom{3}{2} c_{ABC,i} = 3c_{ABC,i}$, with $c_{ABC,i} = 2/\theta_{ABC,i}$. Under our model, at time τ_2, the lineage c from species C must be in subpopulation 2, whereas lineages a and b from species A and B, or the coalesced lineage (ab), could be in either subpopulation. Above the root, once the first coalescence has occurred, the generic instantaneous rate matrix $\mathbf{Q}_{XY}$ models the dynamics. Letting X be a coalesced pair of lineages and Y be the remaining lineage, we have

$$\mathbf{Q}_{XY} := \begin{array}{c} \\ x_1y_2 \\ x_2y_1 \\ x_1y_1 \\ x_2y_2 \\ (xy)_1 \\ (xy)_2 \end{array} \begin{array}{c} \begin{array}{cccccc} x_1y_2 & x_2y_1 & x_1y_1 & x_2y_2 & (xy)_1 & (xy)_2 \end{array} \\ \left[\begin{array}{cccccc} - & 0 & M_{ABC,1} & M_{ABC,2} & 0 & 0 \\ 0 & - & M_{ABC,1} & M_{ABC,2} & 0 & 0 \\ M_{ABC,2} & M_{ABC,2} & - & 0 & c_{ABC,1} & 0 \\ M_{ABC,1} & M_{ABC,1} & 0 & - & 0 & c_{ABC,2} \\ 0 & 0 & 0 & 0 & - & M_{ABC,2} \\ 0 & 0 & 0 & 0 & M_{ABC,1} & - \end{array}\right] \end{array}, \tag{2}$$

where x represents a lineage from coalesced pair X and y represents the lineage from species Y, and the same assumptions as for matrix $\mathbf{Q}_{\mathrm{AB}}$ hold. These states are similarly referred to as states 1 through 6 (i.e., x_1y_2 through $(xy)_2$) of $\mathbf{Q}_{\mathrm{XY}}$.

Probability Distributions of Gene Tree Topologies under the Model

We treat the dynamics in our model as a continuous-time Markov process that describes the waiting times between events. It follows that the waiting time to a given event is distributed exponentially with rate matrix Q, where events are independent as a consequence of the memoryless property of the exponential distribution (Ross 2014, Chapter 5).

Let T_1 and T_2 be random variables denoting the times to the first and second coalescence going backward in time, respectively. Further, let G and σ be random variables denoting the gene tree and species tree topologies, respectively. With the set of parameters $\boldsymbol{\Omega} = \{\sigma, \boldsymbol{\tau}, \mathsf{M}, \boldsymbol{\theta}\}$, let $f_1(t_1, t_2, g|\boldsymbol{\Omega})$ be the joint density of gene tree $G = g$ and coalescence times $T_1 = t_1$ and $T_2 = t_2$ when $0 \leq \tau_1 \leq t_1 < \tau_2 \leq t_2$. Further, let $f_j(t_1, {}_{t2}, g_j|\boldsymbol{\Omega})$, $j \in \{2, 3, 4\}$, be the joint density of gene tree $G = g_j$ and coalescence times $T_1 = t_1$ and $T_2 = t_2$ when $0 \leq \tau_1 \leq \tau_2 \leq t_1 < t_2$, where $g_2 = ((ab)c)$, $g_3 = ((bc)a)$, and $g_4 = ((ac)b)$. Note that $f_1(t_1, t_2, g|\boldsymbol{\Omega})$ does not accumulate mass for $G \neq ((ab)c)$ because only lineages from species A and B can coalesce along the internal branch $[\tau_1, \tau_2)$.

We can split the joint probability densities into a product of marginal densities $h_{\mathrm{AB},ij}$ and $h_{\mathrm{ABC},ij}$ of going from i lineages to j lineages in the internal branch and the branch above the root, respectively, because the waiting times to coalescence events are independent. These marginal densities describe each possible coalescent history through all of the coalescent states described by instantaneous rate matrices $\mathbf{Q}_{\mathrm{AB}}$, $\mathbf{Q}_{\mathrm{XY}}$, and $\mathbf{Q}_{\mathrm{ABC}}$. In order to cohesively describe the different coalescent histories, we introduce mapping functions ϕ_1, ϕ_2, and ϕ_3 that map coalescent states between, respectively, $\mathbf{Q}_{\mathrm{AB}}$ and $\mathbf{Q}_{\mathrm{XY}}$, $\mathbf{Q}_{\mathrm{XY}}$ and $\mathbf{Q}_{\mathrm{ABC}}$, and $\mathbf{Q}_{\mathrm{AB}}$ and $\mathbf{Q}_{\mathrm{ABC}}$, as follows:

$$\phi_1(s) = \begin{cases} 1 & \text{if } s = 5 \\ 4 & \text{if } s = 6 \end{cases}, \tag{3}$$

$$\phi_2(s) = \begin{cases} 5 & \text{if } s = 1 \\ 6 & \text{if } s = 2 \\ 2 & \text{if } s = 3 \\ 8 & \text{if } s = 4 \end{cases}, \tag{4}$$

$$\phi_3(s) = \begin{cases} 3 & \text{if } s \in \{9, 13, 17\} \\ 1 & \text{if } s \in \{10, 14, 18\} \\ 2 & \text{if } s \in \{11, 15, 19\} \\ 4 & \text{if } s \in \{12, 16, 20\} \end{cases}. \tag{5}$$

The function $\phi_1(s)$ maps a coalesced state s in matrix $\mathbf{Q}_{\mathrm{AB}}$ to a corresponding state in matrix $\mathbf{Q}_{XY}$ (e.g., mapping state 5 in $\mathbf{Q}_{\mathrm{AB}}$, where lineages a and b are coalesced in subpopulation 1, into corresponding state 1 of $\mathbf{Q}_{XY}$, where the coalesced lineage is in subpopulation 1 and lineage c is in subpopulation 2). The function $\phi_2(s)$ maps uncoalesced state s in matrix $\mathbf{Q}_{\mathrm{AB}}$ to a corresponding state in matrix $\mathbf{Q}_{\mathrm{ABC}}$ (e.g., mapping state 4 in $\mathbf{Q}_{\mathrm{AB}}$, where lineages a and b are both in subpopulation 2, to state 8 in $\mathbf{Q}_{\mathrm{ABC}}$, where all uncoalesced lineages are in subpopulation 2). The function $\phi_3(s)$ maps a coalesced state s in matrix $\mathbf{Q}_{\mathrm{ABC}}$ to a corresponding state in matrix $\mathbf{Q}_{XY}$ (e.g., mapping state 19 in $\mathbf{Q}_{\mathrm{ABC}}$, where lineages b and c are coalesced in subpopulation 2 and lineage a is in subpopulation 1, to state 2 in $\mathbf{Q}_{XY}$, where the coalesced lineage is in subpopulation 2 whereas the remaining lineage is in subpopulation 1). As one example, to capture the path to a common ancestor between the three lineages drawn in the top left of Figure 1*b*, we see that the system is in state 5 of $\mathbf{Q}_{\mathrm{AB}}$ $[(ab)_1]$ before speciation time τ_2. Therefore at this time, $\phi_1(5)$ is invoked to capture that lineages a and b coalesced along the internal branch in subpopulation 1, whereas the remaining lineage is also in subpopulation 2. This mapping takes us from state 5 of $\mathbf{Q}_{\mathrm{AB}}$ to state 1 of $\mathbf{Q}_{\mathrm{XY}}$ $[(x_1y_2)]$, where here a and b are represented by coalesced lineage x, and c is represented by lineage y.

Put symbolically, let $c_{\mathrm{AB},s}$ and $c_{\mathrm{ABC},s}$ be the coalescence rates leading to state s. For example, if the system has n lineages in subpopulation i, and a coalescence is about to occur there, then the rate of coalescence is $c_{\mathrm{AB},s} = \binom{n}{2} c_{\mathrm{AB},i}$ for the internal branch and $c_{\mathrm{ABC},s} = \binom{n}{2} c_{\mathrm{ABC},i}$ above the root. The marginal densities that describe the possible coalescent histories are

$$\begin{aligned} f_1(t_1,t_2,g|\boldsymbol{\Omega}) = {} & \sum_{s_1=5}^{6} h_{\mathrm{AB},21}(t_1-\tau_1, s_1|\boldsymbol{\Omega}) \\ & \times \left[\sum_{s_2=5}^{6} (\mathrm{e}^{\mathbf{Q}_{\mathrm{AB}}(\tau_2-t_1)})_{s_1 s_2} h_{\mathrm{ABC},21}(t_2-\tau_2, \phi_1(s_2)|\boldsymbol{\Omega}) \right], \\ & \times \mathbf{1}_{\{g=((ab)c) \text{ and } 0\leq\tau_1\leq t_1<\tau_2\leq t_2\}} \end{aligned} \tag{6}$$

$$
\begin{aligned}
f_2(t_1,t_2,g|\boldsymbol{\Omega}) = \sum_{s_1=1}^{4} & h_{AB,22}(\tau_2-\tau_1,s_1|\boldsymbol{\Omega}) \\
& \times \left[\sum_{s_2=9}^{12} h_{ABC,32}(t_1-\tau_2,\phi_2(s_1),s_2|\boldsymbol{\Omega}) \right. \\
& \left. \times h_{ABC,21}(t_2-t_1,\phi_3(s_2)|\boldsymbol{\Omega}) \right] \mathbf{1}_{\{g=((ab)c)} \\
& \quad \text{and } 0 \le \tau_2 \le t_1 < t_2 \}, \\
f_3(t_1,t_2,g|\boldsymbol{\Omega}) = \sum_{s_1=1}^{4} & h_{AB,22}(\tau_2-\tau_1,s_1|\boldsymbol{\Omega}) \\
& \times \left[\sum_{s_2=13}^{16} h_{ABC,32}(t_1-\tau_2,\phi_2(s_1),s_2|\boldsymbol{\Omega}) \right. \\
& \left. \times h_{ABC,21}(t_2-t_1,\phi_3(s_2)|\boldsymbol{\Omega}) \right] \mathbf{1}_{\{g=((ac)b)} \\
& \quad \text{and } 0 \le \tau_2 \le t_1 < t_2 \}, \\
f_4(t_1,t_2,g|\boldsymbol{\Omega}) = \sum_{s_1=1}^{4} & h_{AB,22}(\tau_2-\tau_1,s_1|\boldsymbol{\Omega}) \\
& \times \left[\sum_{s_2=17}^{20} h_{ABC,32}(t_1-\tau_2,\phi_2(s_1),s_2|\boldsymbol{\Omega}) \right. \\
& \left. \times h_{ABC,21}(t_2-t_1,\phi_3(s_2)|\boldsymbol{\Omega}) \right] \mathbf{1}_{\{g=((bc)a)} \\
& \quad \text{and } 0 \le \tau_2 \le t_1 < t_2 \}, \qquad (9)
\end{aligned}
$$

where

$$
\begin{aligned}
h_{AB,21}(t,s|\boldsymbol{\Omega}) &= c_{AB,s}(e^{Q_{AB}t})_{1s} \\
h_{AB,22}(t,s|\boldsymbol{\Omega}) &= (e^{Q_{AB}t})_{1s} \\
h_{ABC,21}(t,s|\boldsymbol{\Omega}) &= c_{ABC,1}(e^{Q_{ABC}t})_{s5} + c_{ABC,2}(e^{Q_{ABC}t})_{s6}. \\
h_{ABC,32}(t,s_1,s_2|\boldsymbol{\Omega}). &= c_{ABC,s}(e^{Q_{ABC}t})_{s_1s_2}.
\end{aligned} \qquad (10)
$$

For $h_{AB,21}(t,s|\boldsymbol{\Omega})$ and $h_{AB,22}(t,s|\boldsymbol{\Omega})$, s represents the state of the system, according to instantaneous rate matrix $\mathbf{Q}_{AB}$, at time t. Specifically, $s \in \{5,6\}$ for $h_{AB,21}$ and $s \in \{1,2,3,4\}$ for $h_{AB,22}$. For $h_{ABC,32}(t,s_1,s_2|\boldsymbol{\Omega})$, s_1 and s_2 represent the states at time 0 and at time t, respectively. That is, $s_1 \in \{5,6,7,8\}$ and $s_2 \in \{9,10,\ldots,20\}$. Similarly, for $h_{ABC,21}(t,s|\boldsymbol{\Omega})$, s represents the state of the system at $t=0$, with $s \in \{1,2,3,4\}$. Though long, equations (6)–(9) are simply capturing the probability density of all coalescent histories (fig. 1*b*).

It follows that the probability of observing gene tree topology G given the model parameters $\boldsymbol{\Omega}$ can be computed from these densities as:

$$
\begin{aligned}
\mathbb{P}[G=((ab)c)|\boldsymbol{\Omega}] = & \int_{\tau_2}^{\infty}\int_{\tau_1}^{\tau_2} f_1(t_1,t_2,((ab)c)|\boldsymbol{\Omega})dt_1dt_2 \\
& + \int_{\tau_2}^{\infty}\int_{\tau_2}^{t_2} f_2(t_1,t_2,((ab)c)|\boldsymbol{\Omega})dt_1dt_2,
\end{aligned} \qquad (11)
$$

$$
\mathbb{P}[G=((ac)b)|\boldsymbol{\Omega}] = \int_{\tau_2}^{\infty}\int_{\tau_2}^{t_2} f_3(t_1,t_2,((ac)b)|\boldsymbol{\Omega})dt_1dt_2, \qquad (12)
$$

$$
\mathbb{P}[G=((bc)a)|\boldsymbol{\Omega}] = \int_{\tau_2}^{\infty}\int_{\tau_2}^{t_2} f_4(t_1,t_2,((bc)a)|\boldsymbol{\Omega})dt_1dt_2. \qquad (13)
$$

Maximum Likelihood Parameter Estimation

It is possible to compute the likelihood from a set of alignments $\{A_1,A_2,\ldots,A_K\}$ at K independent loci as

$$
L(\boldsymbol{\Omega};A_1,A_2,\ldots,A_K) = \prod_{i=1}^{K}\int_G \mathbb{P}[A_i|G]\cdot f(G|\boldsymbol{\Omega})dG, \qquad (14)
$$

where $\mathbb{P}[A_i|G]$ is the probability of observing the ith alignment A_i given gene tree G. The reason that we integrate over gene tree G rather than sum over gene trees is that a gene tree in this context represents a topology together with branch lengths, where the distribution over branch lengths is continuous. Although the likelihood is tractable, it is computationally challenging and changes with the choice of substitution model used to compute the term $\mathbb{P}[A_i|G]$ (Hey and Nielsen 2007). As such, it is common practice in likelihood methods to use estimated gene trees, $\widehat{g}_1,\widehat{g}_2,\ldots,\widehat{g}_K$, as input data, and treat these data as known. Assuming the collection of K inferred gene trees are independent, the likelihood of the model may be computed as

$$
L(\boldsymbol{\Omega};\widehat{g}_1,\widehat{g}_2,\ldots,\widehat{g}_K) = \prod_{i=1}^{K} f(\widehat{g}_i|\boldsymbol{\Omega}). \qquad (15)
$$

Denoting $\boldsymbol{\Omega}_{-\sigma} = \boldsymbol{\Omega}\setminus\{\sigma\}$ as the parameter set $\boldsymbol{\Omega}$ excluding the species tree topology σ, then for a fixed configuration of species tree topology σ, we search over all possible values of unknown parameters in $\boldsymbol{\Omega}_{-}\sigma$ to maximize the log likelihood. Thus, each configuration of the species tree topology is linked with its own set of maximum likelihood estimates $\widehat{\boldsymbol{\Omega}}_{-\sigma}$. We infer the species tree and associated parameter estimates $\widehat{\boldsymbol{\Omega}}$ that give rise to the largest associated likelihood.

It must be emphasized that in this scenario of ancestral structure with two subpopulations, although the three possible species tree topologies are ((AB)C), ((AC)B), and ((BC)A), there are six total species tree model configurations in which species branch from ancestral subpopulations. The reason is that for each species tree topology $((XY)Z)$ relating species X, Y, and Z, it is possible that either Y and Z or X and Z descend from the same ancestral subpopulation. Therefore, we must find the maximum likelihood parameter estimates for six configurations of the species tree.

Extension to *n* Taxa

With three species, there are only three species tree topologies and two configurations of each topology to consider.

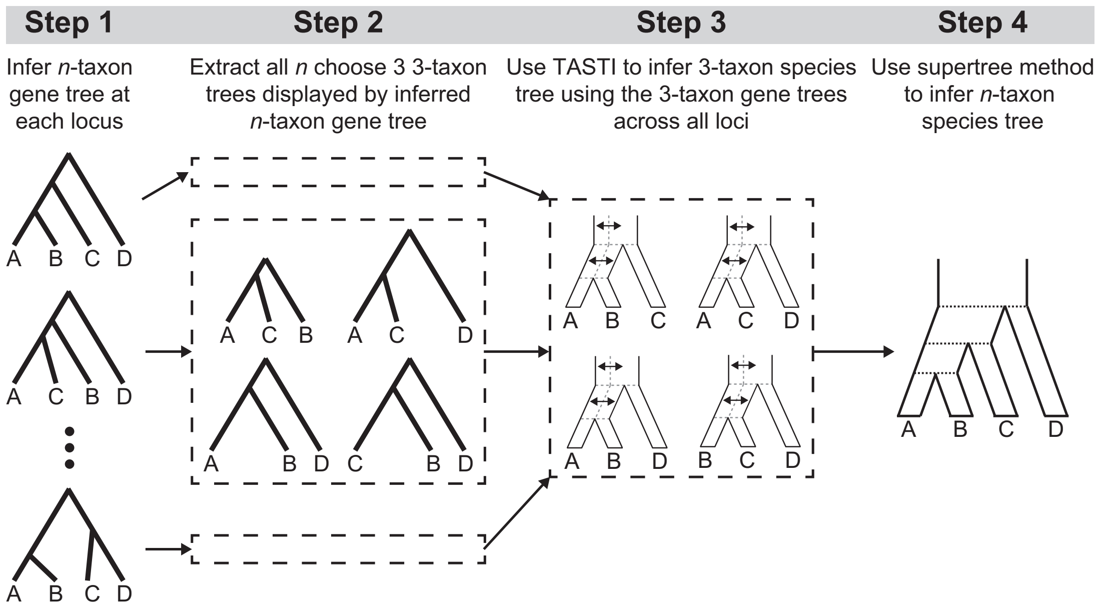
Step 1
Infer *n*-taxon gene tree at each locus
Step 2
Extract all *n* choose 3 3-taxon trees displayed by inferred *n*-taxon gene tree
Step 3
Use TASTI to infer 3-taxon species tree using the 3-taxon gene trees across all loci
Step 4
Use supertree method to infer *n*-taxon species tree
A B C D
A C B D
A B C D
A C B
A C D
A B D
C B D
A B C
A C D
A B D
B C D
A B C D

Fig. 2—Schematic of a supertree approach for inferring *n*-taxon species tree topologies under a model of ancestral population structure. (Step 1) At each locus, infer a rooted *n*-taxon (here $n = 4$) gene tree. (Step 2) For each gene tree, extract the set of $\binom{n}{3}$ (here equaling four) rooted three-taxon gene trees compatible with each *n*-taxon gene tree. (Step 3) For a given set of three taxa (e.g., species A, B, and C), apply TASTI to infer a rooted three-taxon species tree under a model of ancestral population structure by using all three-taxon gene trees with the given set of taxa across all loci. (Step 4) Given the set of $\binom{n}{3}$ (here equaling four) rooted three-taxon species trees inferred by TASTI, use a supertree approach to infer a rooted *n*-taxon species tree topology for the full set of *n* taxa.

However, as the number of taxa grows, the number of available topologies, which grows in complexity at a rate proportional to the double factorial of the number of taxa n (Felsenstein 2004), makes direct application of our method computationally infeasible. However, as illustrated in figure 2, dividing the *n*-taxon problem into multiple subproblems can ameliorate this burden. Because a rooted *n*-taxon tree is characterized by its set of rooted triples (Steel 1992), we can implement a supertree approach (DeGiorgio and Degnan 2010) to estimate the *n*-taxon species tree topology. This tactic reduces computation time to the order of $\binom{n}{3}$ times the original complexity of the problem on three taxa.

Implementation

TASTI jointly optimizes over divergence times and migration rates, though the computation varies depending on the type of input data—that is, gene tree topologies with, or without, branch lengths. From this point forward, for simplicity and clarity, we drop the “hats” on our estimates for gene trees, as we treat them as fully known.

We first consider the case when branch lengths are included to make inference. Under this scenario, estimated coalescence times t_1 and t_2 are already fixed, eliminating the need to integrate out their values and permitting direct computation of the probability density given $\boldsymbol{\Omega}$. For coalescence times t_{1i} and t_{2i} of gene tree g_i, computing the likelihood reduces to

$$L(\boldsymbol{\Omega}\,; g_1, g_2, \dots g_K) = \prod_{i=1}^{K}\sum_{j=1}^{4} f_j(t_{1i}, t_{2i}, g_i|\boldsymbol{\Omega}). \qquad (16)$$

On the other hand, when simply inputting gene tree topologies, we can decrease the size of the parameter space by optimizing over the length of the internal branch, rather than estimating each divergence time separately, because our model assumes no coalescence occurs in the interval $[0, \tau_1)$. The problem posed in equations (11)–(13) also reduces to integrating only over t_1, as the coalescence event at t_1 uniquely determines the topology of the tree. Given data $\mathbf{y} = (n_{ab}, n_{ac}, n_{bc})$, where n_{xy} represents the number of three-taxon gene tree topologies in a sample displaying clade $\{x, y\}$ for lineages x and y from species X and Y, we can compute the likelihood under $\boldsymbol{\Omega}$ as

$$\begin{aligned} L(\boldsymbol{\Omega};\, \mathbf{y}) = {} & \mathbb{P}[G = ((ab)c)|\boldsymbol{\Omega}]^{n_{ab}} \times \mathbb{P}[G = ((bc)a)|\boldsymbol{\Omega}]^{n_{bc}} \\ & \times \mathbb{P}[G = ((ac)b)|\boldsymbol{\Omega}]^{n_{ac}}. \end{aligned} \qquad (17)$$

To implement a constrained optimization procedure, we first need to derive bounds on our unknowns. First, we consider the migration rate. Recall that the migration rate is given by $M = 4Nm/\theta$, where $m \in (0, 1]$. This formulation gives rise to a lower bound of $M > 0$ when m is arbitrarily close to 0, since $m = 0$ would imply no migration between subpopulations and thus would prevent the process from stopping. Similarly, we have an upper bound on M of $4N/\theta$ when $m = 1$. As the migration rate grows toward its upper bound, the species ancestry becomes effectively unstructured, generalizing our model to the standard multispecies coalescent.

If branch lengths are used as input, then the estimated coalescence times determine the bounds on the speciation times. Let $\{T_{11}, T_{12}, \ldots, T_{1K}\}$ and $\{T_{21}, T_{22}, \ldots, T_{2K}\}$ be the collections of first and second coalescent times at K loci, respectively, going backward in time. Then by our model construction, $\tau_1 \in [0, \min\{T_{11}, T_{12}, \ldots, T_{1K}\}]$ and $\tau_2 \in [\tau_1, \min\{T_{21}, T_{22}, \ldots . T_{2K}\}]$. In a maximum likelihood framework, if branches of a gene tree in the sample are uninformative, it is then possible to encounter a "star tree" scenario in which $\tau_1 = \tau_2 = 0$. Here, we cannot resolve the relationship among the taxa and are left to infer the unresolved species tree topology (ABC). Additional details regarding the implementation of the optimization procedure are discussed in the supplementary section "Parameter Estimation for Gene Trees with Branch Lengths," Supplementary Material online.

When only topologies are considered, however, the internal branch length is no longer bounded by the set of observed divergence times. Therefore, we bound the length of the internal branch of the species tree in a manner similar in spirit to other methods that assume some minimum level of incomplete lineage sorting (Than et al. 2008; Wu 2012). To do this, we claim it is reasonable to search only over internal branch lengths that allow for some reasonable level of gene tree discordance (Hudson 1983; Tajima 1983; Pamilo and Nei 1988). This is done because as the length of the internal branch of the species tree increases, the probability of discordance in the presence of population structure, even with low migration rates, goes to zero. Letting τ represent the length of the internal branch, this probability of discordance is computed by

$$\mathbb{P}[G \neq \sigma] = \frac{2}{3} e^{-2\tau/\theta}. \tag{18}$$

We propose that the lowest reasonable level of discordance to assume is $\mathbb{P}[G \neq \sigma] = 0.05$, implying that out of 100 sampled loci, only five will exhibit discordance.

With topology data only, an analysis of three taxa with TASTI takes approximately 10 min, regardless of the number of observed loci. When including branch lengths, the running time grows with the number of observed loci. As the number n of taxa grows, complexity is multiplied by the number $\binom{n}{3}$ of rooted triples, but the computing task is trivially parallelizable across these triples (see supplementary fig. S2, Supplementary Material online).

Results

In this section, we aim to examine the performance and robustness of our proposed method using various forms of input data and compare this performance with existing methods MP-EST (Liu, Yu, and Edwards 2010), STELLS2 (Pei and Wu 2017), and STEM (Kubatko et al. 2009). MP-EST is a pseudo-likelihood approach based on triples of taxa, whereas STELLS2 and STEM are likelihood approaches based on gene tree topologies and gene tree topologies with branch lengths, respectively. We selected these methods for comparison specifically because they are state-of-the-art methods for estimating species trees from gene trees which 1) operate within the maximum likelihood paradigm but 2) do not assume any sort of structure or other inter- or intraspecies gene flow. For TASTI, we considered using gene tree topologies with and without branch lengths in cases where the input data are both inferred and known with certainty. The gene trees in our simulations were generated using the coalescent simulator *ms* (Hudson 2002). To simulate sequence alignments conditional on these gene trees, we employed Seq-Gen (Rambaut and Grass 1997) to generate 1-kilobase (kb) and 0.5-kb-long sequences (with an additional outgroup sequence) under the HKY substitution model (Hasegawa et al. 1985) with a transition–transversion ratio of 4.6 and base frequencies A, T, C, and G, respectively, equal to 0.3, 0.3, 0.2, and 0.2.

From these simulated sequences, we used dnamlk of PHYLIP under the HKY substitution model (Felsenstein 1989) to infer gene trees using maximum likelihood assuming a molecular clock with a transition–transversion ratio of 4.6 and empirical base frequencies. We applied this pipeline to create samples of 100 replicates of K independent loci, where K ranged from 10 to 10^4, under fixed species tree topology ((AB)C). We let $M_{\mathrm{AB},1} = M_{\mathrm{AB},2} = M_{\mathrm{ABC},1} = M_{\mathrm{ABC},2} = M$ and $\theta_{\mathrm{AB},1} = \theta_{\mathrm{AB},2} = \theta_{\mathrm{ABC},1} = \theta_{\mathrm{ABC},2} = \theta$. We set a constant effective population size of $N = 5 \times 10^4$ across both subpopulations and employed a per-site per-generation mutation rate of $\mu = 2.5 \times 10^{-8}$, yielding $\theta = 0.005$. When gene tree topologies were used as input, this resulted in a bound on the length of the internal branch of the species tree of $\tau_2 - \tau_1 \approx 6.5 \times 10^{-3}$ mutation units, as derived from equation (18). We focus on this case, with a short internal branch, as it is a challenging scenario which helps elucidate method performance under the most extreme settings. We do, however, further evaluate the performance of TASTI in cases with longer internal branch lengths (see, e.g., table 1, fig. 4, or supplementary fig. S9, Supplementary Material online). These parameter settings were inspired by the great ape data set of Burgess and Yang (2008).

Table 1

Accuracy of the Maximum Likelihood Estimator as a Function of Migration Rate $M = 4Nm/\theta$

	Exact Topologies			Inferred Topologies			Exact Gene Trees			Inferred Gene Trees		
M	**((AB)C)**	**((BC)A)**	**((AC)B)**	**((AB)C)**	**((BC)A)**	**((AC)B)**	**((AB)C)**	**((BC)A)**	**((AC)B)**	**((AB)C)**	**((BC)A)**	**((AC)B)**
0.5	60	23	17	63	19	18	84	0	16	75	0	25
5	73	27	0	75	25	0	94	0	6	93	0	7
50	77	23	0	75	25	0	94	0	6	91	0	9
$4N/\theta$	74	13	13	52	25	23	100	0	0	84	5	11

NOTE.—Accuracy is based on the percentage of 100 simulated replicates of 10^3 loci that inferred a specific species tree topology under a scenario with $\tau_1 = 2.5 \times 10^{-3}$ and $\tau_2 = 5.25 \times 10^{-3}$.

Ancestral Population Structure Alters Expected Gene Tree Distributions

Although symmetries among the distribution of gene tree topologies are expected under the standard multispecies coalescent (Allman et al. 2011), population structure skews this distribution, as lineages above the root are no longer guaranteed to have equal probability of coalescing (Slatkin and Pollack 2008). In supplementary figure S3, Supplementary Material online, distributions of gene trees are plotted for varying levels of migration when the internal branch is 6.5 $\times 10^{-3}$ mutation units. Indeed, as $M \to 4N/\theta$—the scenario under which our model reduces to the standard multispecies coalescent—this symmetry between topologies $((bc)a)$ and $((ac)b)$ is present. However, as the migration rate decreases, the distribution skews toward a dominant $((bc)a)$ topology, and symmetry between gene tree topologies $((bc)a)$ and $((ac)b)$ no longer persists. Further, although estimated gene tree topologies (supplementary fig. S3*c*, Supplementary Material online) obey the same large-sample distribution as the topologies known exactly (supplementary fig. S3*a*, Supplementary Material online), there is an increased variability in their distribution across samples, which in general is detrimental to inference (Casella and Berger 2002, Chapter 10).

TASTI incorporates the possibility for these skewed distributions of gene tree topologies. In what follows, we demonstrate that TASTI provides reasonable estimates of the true species tree topology under conditions in which other methods have previously failed (DeGiorgio and Rosenberg 2016). We simulated data as described at the start of this section, assuming a species tree with speciation times $\tau_1 = 2.5 \times 10^{-3}$ and $\tau_2 = 2.75 \times 10^{-3}$ mutation units. This setting, with such a short internal branch, makes correct inference especially challenging. In order to understand the effects of various evolutionary parameters on TASTI's performance, we first conduct a detailed exploration of the method's accuracy. A summary of TASTI's accuracy is in figure 3, whereas tables of median parameter estimates and 95% confidence intervals are in supplementary tables S1 and S2, Supplementary Material online, respectively. Summaries of additional simulations studying the performance of TASTI against MP-EST, STELLS2, and STEM under a variety of different simulation settings are in figure 4.

Our results drive home several key points. First, phylogenetic inference methods are sensitive to the quality of input data, and TASTI taking gene tree topologies as input tends to be more robust to stochasticity encountered in practice than when gene tree branch length information is incorporated into TASTI's input as well. However, if accurate data on gene tree branch lengths are available, then TASTI will naturally perform better when this information is included rather than obscured. This is especially true when multiple candidate species trees are equally likely under a model of gene tree topologies alone. Furthermore, when species trees have longer internal branch lengths, then more accurate input data can be estimated, yielding more reliable downstream inference of the complete species trees. Lastly, although TASTI is typically competitive with alternative methods under the standard multispecies coalescent, it is the only method that exhibits favorable performance in the presence of ancestral population structure.

Gene Tree Topology Data Permit Accurate, Robust Species Tree Inference

The first column of figure 3 shows the performance of TASTI with exact gene tree topologies for data input across all investigated levels of migration. In the case of no ancestral structure ($M = 4N/\theta$), we obtain consistent estimates of the true species tree even with a small number of sampled loci. Similarly, when the ancestral populations are structured with $M = 5$, TASTI's species tree estimate begins favoring the truth over alternative topologies fairly quickly. In contrast, with a combination of a higher rate of migration between subpopulations ($M = 50$) and a short internal branch of the species tree, the effects of ancestral population structure begin to break down. Consequently, the distribution of gene tree topologies resembles one we would expect under the standard multispecies coalescent with sister species B and C, such that TASTI is led to infer the dominant gene tree topology ((BC)A) as the true species tree. When migration is at its lowest ($M = 0.5$), and especially with few sampled loci, the

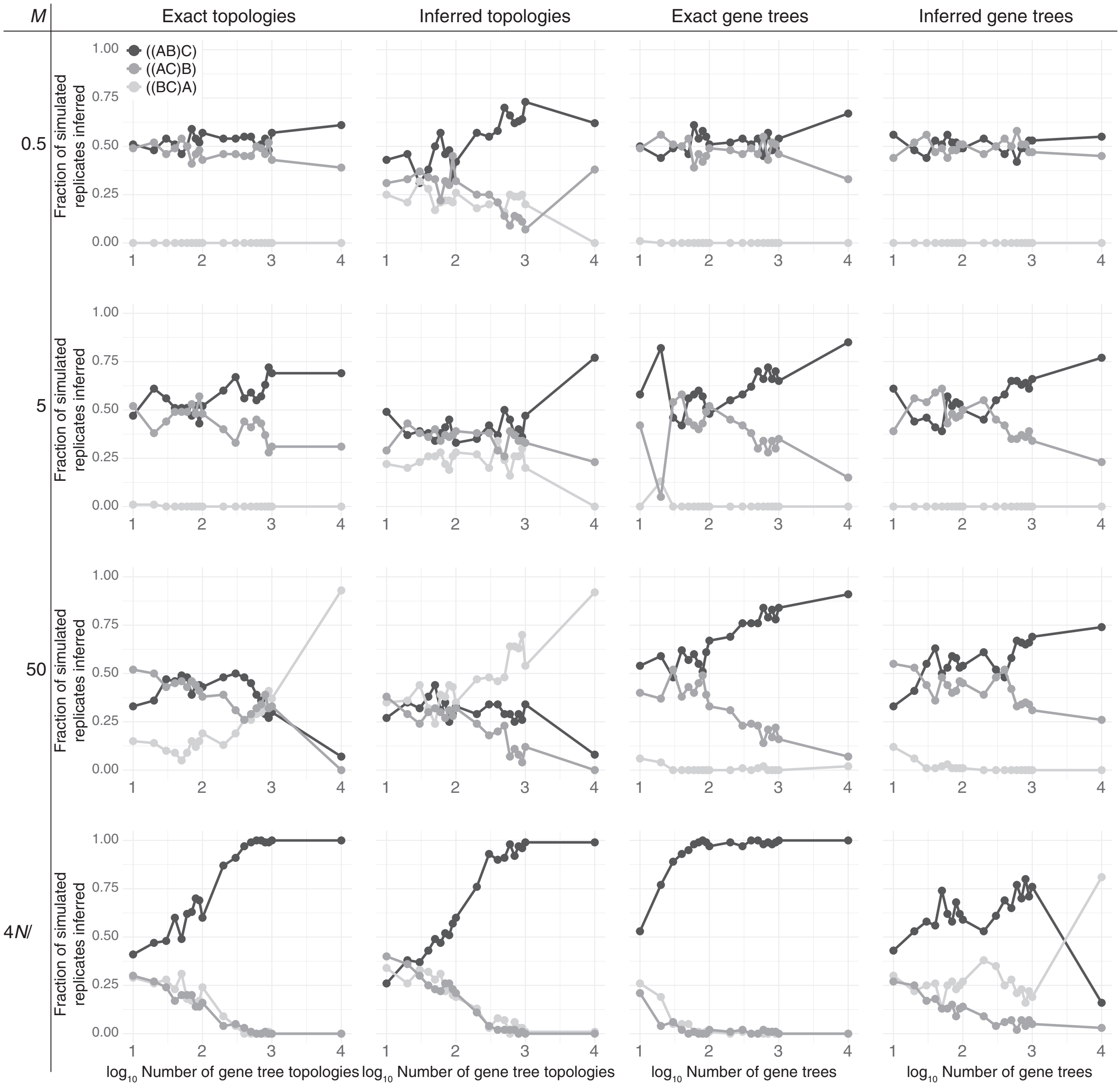

Fig. 3.—Accuracy of the maximum likelihood estimator of σ as a function of the number of input gene trees. Accuracy is based on the proportion of 100 simulated replicates of K loci (before filtering "star trees"), with K ranging from 10 to 10^4, where our method inferred a specific species tree topology under a scenario with $\tau_1 = 2.5 \times 10^{-3}$ and $\tau_2 = 2.75 \times 10^{-3}$.

data do not convey any detectable signal in the distribution of topologies. That is, nearly all of the observed topologies are of the form $((bc)a)$, and the true species tree is difficult to determine. Interestingly, though $((bc)a)$ is by far the dominant gene tree in these input data, ((BC)A) is never TASTI's estimated maximum likelihood species tree. This result is expounded upon in supplementary section "Bounding the Length of the Internal Branch of the Species Tree in the Low Signal Setting," Supplementary Material online.

The second column of figure 3 illustrates TASTI's accuracy when gene tree topologies are no longer known with certainty, but rather, are inferred from 1-kb sequences. A comparison of these results with those in the first column suggests that TASTI is robust to noise incurred in the process of gene tree inference. Although accuracy in general is reduced, the overall trends are still present. For the no structure setting, and when ancestral populations are structured with $M = 5$, TASTI still favors the true species tree topology ((AB)C), but with less

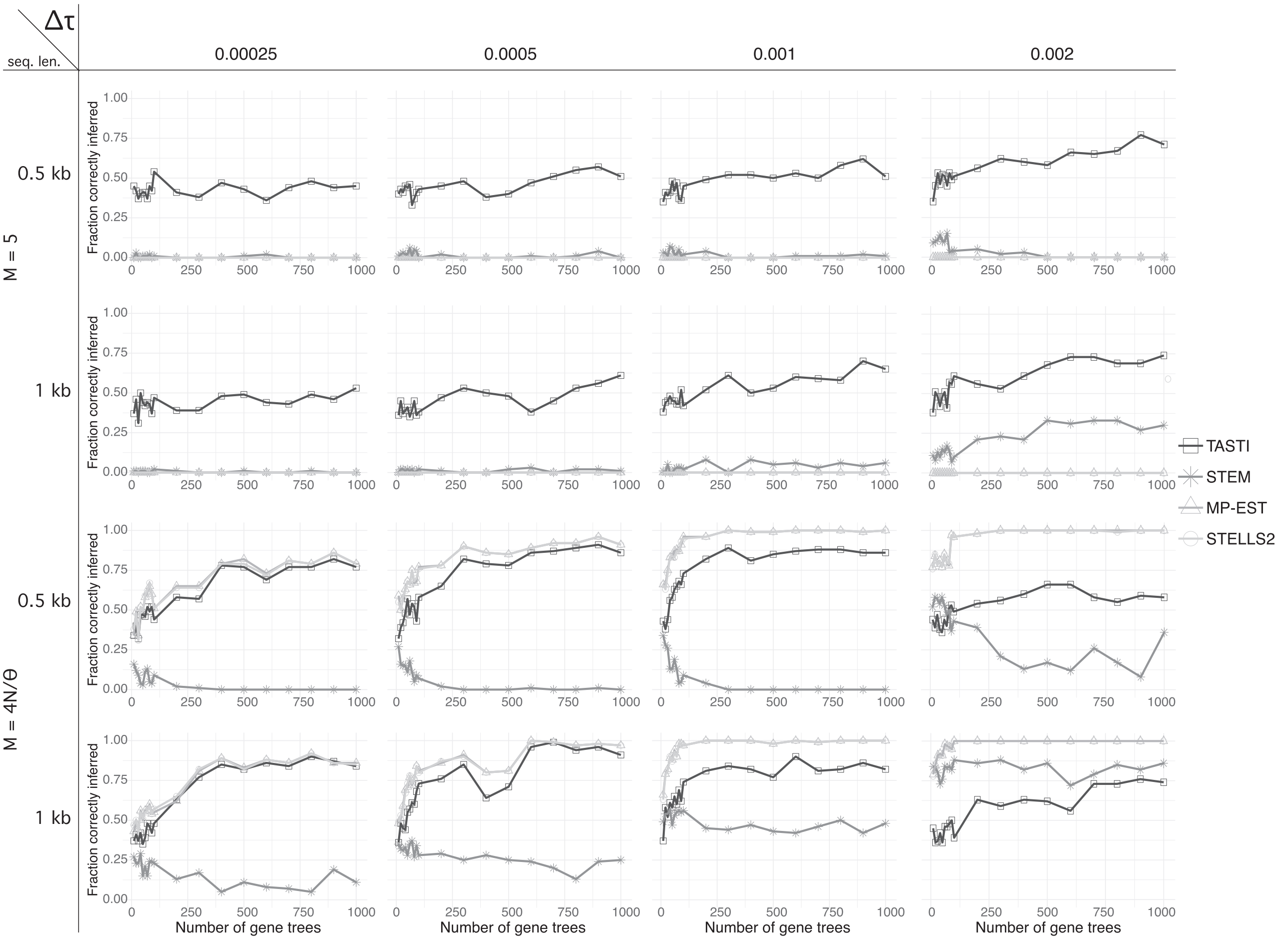

Fig. 4.—Accuracy of TASTI, MP-EST, STELLS2, and STEM when estimating the true species tree topology σ as a function of the number of input gene trees. Accuracy is based on the proportion of 100 simulated replicates of *K* loci with *K* ranging from 10 to 10^3. In these simulations, we only use inferred gene tree topologies as input to TASTI. Comparisons are made across four internal branch lengths ($\Delta\tau = \tau_2 - \tau_1$) in cases when populations are both structured and unstructured and when gene trees are estimated from both 1- and 0.5-kb sequences.

immediacy and certainty as is found in the case when topologies are known exactly. When $M = 50$, TASTI goes from slowly favoring the misleading species tree topology ((BC)A) to quickly being positively misleading for ((BC)A). Interestingly, when migration is at its lowest, the gene tree distribution skews toward a more frequent occurrence of $((ab)c)$ than $((ac)b)$ when compared against the distribution of gene tree topologies which are known exactly. This in turn improves TASTI's accuracy.

Knowledge of Gene Tree Branch Lengths Improves Estimation

Incorporating branch length data from input gene trees can improve inference under TASTI, as knowledge about presence or absence of symmetries in the distribution of gene tree topologies may sometimes be insufficient for making the correct judgment. The performance of TASTI when gene tree topologies with branch lengths are known with certainty, displayed in the third column of figure 3, provides a benchmark best case that TASTI can hope to achieve under our highly challenging simulation settings. In the case when ancestral populations are not structured, we see perfect performance by TASTI with few input loci. With ancestral structure, we still achieve good performance under each level of migration, with weaker trends as structure between subpopulations increases. Remarkably, although TASTI's estimate of the species tree topology is inconsistent when $M = 50$, the addition of branch lengths rescues inference. Another advantage of using branch lengths is that we eliminate the need to set an upper bound on the length of the internal branch heuristically, as the estimated gene trees determine this automatically.

Using estimated gene trees and branch lengths introduces increased variability into the model when compared with the noise introduced by only using estimated topologies. In general, although we obtain fairly accurate species tree estimates using gene tree data known with certainty, the overall trend is that TASTI experiences a larger reduction in accuracy from using estimated gene trees than it does from using estimated topologies alone. For example, we can see that under the standard multispecies coalescent, though we obtained near-perfect performance when branch lengths were known with certainty, the variability in the estimated branch lengths greatly reduces TASTI's accuracy. This trend is apparent across all migration rates, but it increases in severity as population structure decreases. This is consequent to the claim that if gene tree branch lengths are shorter, they are more difficult to estimate accurately (DeGiorgio and Degnan 2014). This notion is explored more deeply in the following section.

Species Trees with Longer Total Branch Length Permit Better Inference

Our simulations test the accuracy of TASTI under a highly challenging scenario where the internal branch of the species tree is excessively short. With a short internal branch under ancestral population structure, the chance of observing gene trees concordant with the species tree decreases. Additionally, for a fixed migration rate and divergence time τ_1, the time to the MRCA among lineages in the phylogeny decreases on average with a decreasing internal branch length. These shorter gene tree branch lengths make accurate gene tree inference more difficult, as loci become less informative because fewer mutations will have occurred among the species (DeGiorgio and Degnan 2014).

The impact of the short gene tree branch lengths can be seen directly by comparing the accuracy of TASTI when applied to gene trees known with certainty as opposed to when it is applied to inferred gene trees (fig. 3, third and fourth columns). When migration is low, gene tree branch lengths are longer, and hence more informative. Thus, in lower-migration settings (e.g., when $M = 0.5$ or $M = 5$), we can infer more accurate gene trees and branch lengths than in higher-migration scenarios. However, when migration is high (e.g., when $M = 50$) or when the ancestral species are unstructured, gene tree branch lengths are much shorter, thereby leading to less accurately inferred divergence times. We can see that the accuracy of TASTI suffers the most when gene tree branch lengths are inferred under the no ancestral structure and high migration scenarios. Meanwhile, species tree inference is not dramatically affected when migration is low.

With this in mind, we tested the influence of noisy gene trees on TASTI's performance in two ways. First, we considered the effect of less accurate input gene trees by following an identical simulation protocol to the one described above, with the exception that we inferred gene trees from 0.5 kb instead of 1-kb-long regions. Results display similar, though slightly worsened, performance when compared with our original simulations (compare fig. 3 and supplementary fig. S4, Supplementary Material online). We also tested the accuracy of TASTI with a longer internal branch of the true species tree, such that the time to the MRCA is on average longer. To evaluate this scenario, we simulated 100 replicates of 10^3 loci using the same parameters and protocols as previously described, with the exception that we increased the age τ_2 of the species tree root. Specifically, we employed divergence times of $\tau_1 = 2.5 \times 10^{-3}$ and $\tau_2 = 5.25 \times 10^{-3}$. As expected, we observe the same trends as in our original simulations (fig. 3) but with overall more accurate species tree inference (table 1). This is still a challenging setting, noting that the probabilities of observing concordant gene trees are 3.62×10^{-3}, 3.46×10^{-2}, 0.236, and 0.596 for $M = 0.5$, 5, 50, and $4N/\theta$, respectively, with the longer internal branch length. This can be compared against concordance probabilities of 1.83×10^{-3}, 1.75×10^{-2}, 0.122, and 0.366 for $M = 0.5$, 5, 50, and $4N/\theta$, respectively, with the shorter internal branch length.

Although our study operated on gene trees inferred by maximum likelihood, the use of Bayesian methods to infer gene trees may alleviate these concerns by improving gene tree estimates and decreasing the chance of estimating gene trees with branches of length zero (DeGiorgio and Degnan 2014). Alternatively, it is possible that imposing a constraint on a minimum value on $\tau_2 - \tau_1$ when using whole gene trees would avoid the deleterious effects of gene trees with poorly estimated branch lengths.

Comparison of Competing Methods in the Three-Taxon Setting

We used simulations to compare TASTI with three other methods for inferring species trees from gene trees: MP-EST, STELLS2, and STEM. We designed the study similarly to the one presented in figure 3, examining the methods' collective performance over varying internal branch lengths and levels of population structure. Specifically, we let $(\tau_2 - \tau_1) \in \{2.5 \times 10^{-4}, 5.0 \times 10^{-3}, 1.0 \times 10^{-3}, 2.0 \times 10^{-3}\}$ mutation units and $M \in \{5, 4N/\theta\}$. We only investigate the more realistic case of estimated gene trees as input, where gene trees were inferred from 0.5 and 1-kb-long sequences. For TASTI, we study the setting where gene tree topologies are used as input. Other details of the simulations, including parameters N and θ, remain the same as in previous simulations.

Results (fig. 4) show the proportion of times out of 100 replicate simulations that each method inferred the correct species tree for the given parameter settings. Comparing between the top two rows, we see that TASTI's accuracy increases with more accurate input topologies and with a longer internal branch, where the signals are stronger. Meanwhile, all other methods nearly always infer the wrong species tree topology nearly (or exactly) 100% of the time, demonstrating that MP-EST, STELLS2, and STEM are not robust to ancestral population structure.

The bottom two rows, on the other hand, evaluate all methods when the populations are unstructured. Here, MP-EST and STELLS2 show the same consistent performance; TASTI is competitive with these two methods. Though STEM is known to perform well when gene trees are known exactly, we see that this method is not robust to empirical data and therefore underperforms, especially when the gene trees are inferred from shorter sequences (Leaché and Rannala 2011; DeGiorgio and Degnan 2014). An exception is that, as the internal branch length grows longer, sometimes a structured species tree topology becomes more likely than the unstructured ((AB)C) topology under TASTI. This occurs when stochasticity in the observed data results in asymmetries in the observed distribution of gene tree topologies. Options that would remedy the unidentifiability in this case are to limit the length of the internal branch using some sort of external information or to include divergence times in the analysis. As these options may not always be possible, we recognize this issue as a drawback of TASTI.

Accuracy of the Supertree Approach Applied to Four Taxa

With the wide availability of multilocus sequence data from a number of species (Johnson et al. 2013; Lin et al. 2014; Yang et al. 2015; Shen et al. 2016), it is common practice to construct species-level phylogenies for more than three taxa. However, when performing model-based species tree estimation, the inference problem becomes increasingly complex and computationally difficult (Drummond and Rambaut 2007; Liu and Pearl 2007; Heled and Drummond 2010; Wu 2012). We therefore chose to evaluate the performance of TASTI for an increased problem size. Using again the identical simulation protocol, we generated data assuming a four-taxon species tree with fixed species tree topology (((AB)C)D) and speciation times $\tau_1 = 1.25 \times 10^{-2}$, $\tau_2 = 1.375 \times 10^{-2}$, and $\tau_3 = 1.5 \times 10^{-2}$ mutation units. Species A originated from subpopulation 1 whereas species B, C, and D originated from subpopulation 2, as depicted in figure 5. We again set a constant effective population size of $N = 5 \times 10^4$ across both of these subpopulations and considered both the unstructured ancestral population case as well as the case with symmetric migration rate of $M = 5$ between subpopulations in ancestral species.

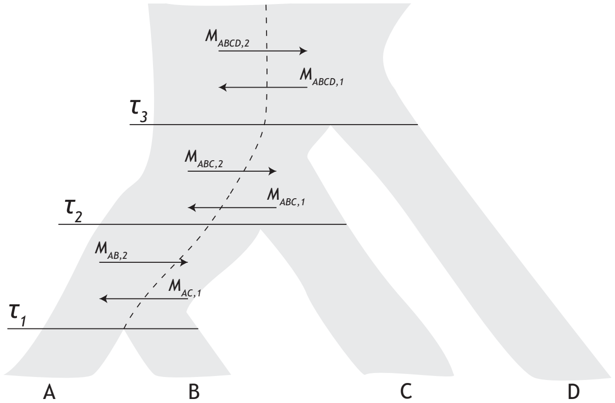

Fig. 5.—Model four-taxon species tree used in our supertree simulations, displaying the relationships among species A, B, C, and D, with divergence times τ_1, τ_2, and τ_3. Ancestral species belong to one of two subpopulations with migration between subpopulations at rates $M_{\mathrm{AB},1}$ and $M_{\mathrm{AB},2}$ directly ancestral to species A and B, $M_{\mathrm{ABC},1}$ and $M_{\mathrm{ABC},2}$ directly ancestral to the split of species A and B with species C, and $M_{\mathrm{ABCD},1}$ and $M_{\mathrm{ABCD},2}$ above the root. Lineages from species A descend from subpopulation 1, whereas lineages from species B, C, and D descend from subpopulation 2. Under our simulation scenarios, we set $M_{X,i} = M_{Y,j} = M$ for all species X and Y and for all subpopulations i and j.

To infer four-taxon species trees, we followed the procedure depicted in figure 2. For each simulated replicate, we extracted the $\binom{4}{3} = 4$ rooted triples displayed by each four-taxon gene tree at each of the K simulated loci. That is, at every locus, we extracted the rooted gene tree (topology with corresponding branch lengths) associated with each set of three species, such that we obtained rooted gene trees for the sets of taxa $\{a, b, c\}$, $\{a, b, d\}$, $\{a, c, d\}$, and $\{b, c, d\}$. For each set of three species, we estimated a three-taxon species tree topology using TASTI based upon the set of rooted triples for that set of three species across the K simulated loci. This procedure yielded $\binom{4}{3} = 4$ three-taxon species tree topology estimates. These four triples were then used to construct a supertree using the modified mincut supertree algorithm (Page 2002).

We chose to use the modified mincut supertree algorithm based on six properties it exhibits (Page 2002; Semple and Steel 2003). First, any species that is in one of the input gene trees is also in the inferred supertree. Second, if a tree exists that could display each input tree as a subtree, then the supertree will display each of these trees. Third, the order in which gene tree topologies are provided to the algorithm does not affect its supertree estimate. Fourth, changing the labeling of species in the set of input trees will produce the same output supertree with the set of relabeled species. Fifth, the algorithm runs in polynomial time as a function of the number of input species n. Finally, nestings that are not contradicted are also displayed in the supertree. Though we employ a specific supertree algorithm here, other supertree approaches may have complementary desirable properties that could lead to improvements in performance. However, an extensive assessment of the wide diversity of supertree algorithms (Wilkinson et al. 2005) is beyond the scope of this article.

In general, the most frequently inferred species tree topologies are the true topology (((AB)C)D) and the partially unresolved topology ((AB)CD) (supplementary fig. S9, Supplementary Material online). These unresolved trees occur as a consequence of the supertree construction algorithm when inferred triples are in conflict. We further investigated the performance of TASTI by examining the

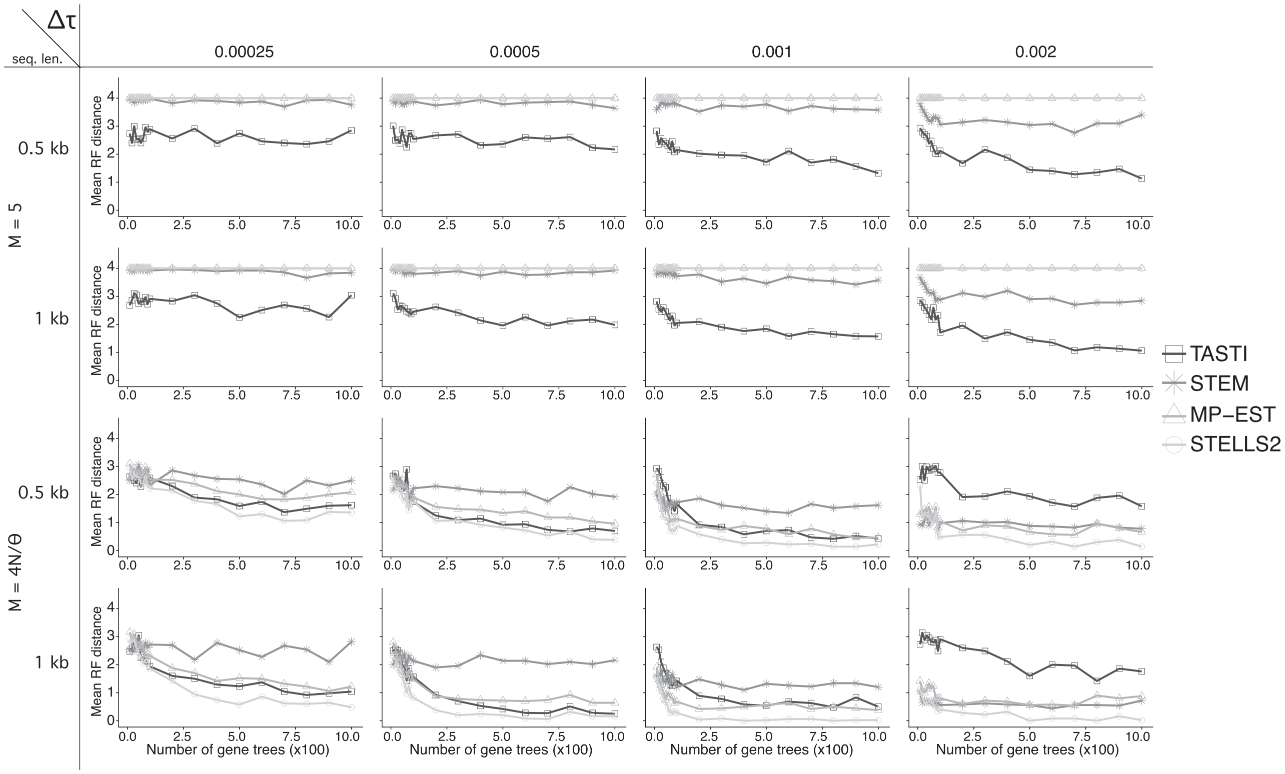

Fig. 6.—Accuracy of the competing methods' estimators of σ as a function of the number of input gene trees in the four-taxon setting. Accuracy is based on the average RF distance between the estimated and true species trees across 100 simulated replicated of K loci, with K ranging from 10 to 10^3. We considered a migration rate $M = 5$ as well as the unstructured scenario.

mean Robinson–Foulds (RF) distance (Robinson and Foulds 1981) between the inferred species tree topology and the true topology for each set of 100 replicates given K simulated loci. We followed the procedure of DeGiorgio and Degnan (2014) and defined the RF distance as the sum of the number of false positive clades and false negative clades between the inferred topology and the true topology, such that the distance metric applies to multifurcating trees. Mean RF distances, numbers of false positive clades, and numbers of false negative clades across all simulations are summarized in supplementary figure S10, Supplementary Material online. Results show that false negative clades occur more frequently than false positive clades. This result indicates that the largest errors of this approach are due to partially unresolved species trees, rather than species trees displaying incorrect clades. Further, RF distance decreases as the number of input loci increases, with the exception of species tree estimates based on inferred gene trees with branch lengths.

Comparing Accuracy of Multiple Methods Applied to Four Taxa

In order to place the above results in better context, we compared TASTI with MP-EST, STELLS2, and STEM in the four-taxon setting as well. We followed the same protocol for simulation that was used in the section "Accuracy with the Supertree Approach Applied to Four Taxa." In each simulation, we assumed equispaced divergence times (e.g., $\tau_3 - \tau_2 = \tau_2 - \tau_1$) along an asymmetric species tree. The simulations include settings where the divergence times are separated by one of 2.5×10^{-4}, 5.0×10^{-4}, 1.0×10^{-3}, or 2.0×10^{-3} mutation units. We considered the structured case with migration rate $M = 5$ and the unstructured case. We applied this comparison only to inferred gene trees, where the gene trees were estimated from 1- and 0.5-kb sequences. For TASTI, we only used gene tree topologies as input.

The results from this simulation, summarized using the RF distance in figure 6, behave as expected (the same simulations, described by number of false positive clades and false negative clades, are reported in supplementary figs. S7 and S8, Supplementary Material online, respectively). The top two rows, showing simulations when the populations are structured, indicate that MP-EST, STELLS2, and STEM behave similarly poorly; each method on average infers a tree that is as far away in RF distance from the true tree as possible. TASTI, on the other hand gets closer to the true tree, in particular as the internal branches grow longer. In the unstructured case, although TASTI, MP-EST and STELLS2 are all competitive, STELLS2 uniformly performs the best. TASTI and MP-EST perform very similarly because MP-EST takes a pseudolikelihood approach that, like TASTI, is also based in rooted triples. Because STEM is sensitive to noisy data, this method again underperforms relative to the other three. As observed in the three-taxon simulations in figure 4, we note one exception where TASTI underperforms when the internal branch length grows but the populations are unstructured. The driver of TASTI's underperformance remains the same, resulting in an increase in false negative clades (supplementary fig. S8, Supplementary Material online), though TASTI performs similarly to MP-EST and STELLS2 with regard to false positive clades (supplementary fig. S7, Supplementary Material online).

Application to Data from the *Anopheles gambiae* Complex

The *Anopheles gambiae* complex comprised several morphologically indistinguishable yet genetically different species of mosquitos living in sympatry in sub-Saharan Africa (Davidson 1962). Though the species belonging to the *A. gambiae* complex are closely related, only a small fraction of them possess significant malaria vectorial capacity. Studying their ancestry is critical as it may lead to key insights regarding how evolutionary history affects the development of traits that drive successful malaria vectors (Neafsey et al. 2015). Fontaine et al. (2015) originally studied six species in this complex: *Anopheles arabiensis*, *Anopheles coluzzii*, *A. gambiae sensu stricto*, *Anopheles melus*, *Anopheles merus*, and *Anopheles quadriannulatus*. In their work, they identified that two of the most important malaria vectors in the complex, *A. arabiensis* and *A. gambiae*, are in fact not the most closely related. However, the similarities in their genome could be attributed to introgression between them found on their autosomes. Wen, Yu, Hahn, et al. (2016) built on the analyses of Fontaine et al. (2015) by applying a phylogenetic network model to the *Anopheles* data.

Previously, however, Lehmann et al. (1998) found significant evidence for population structure within the *A. gambiae* complex. They note that, though anopheline species are broadly present across Africa, there are regions where populations generally cannot establish, thereby isolating some subpopulations. Moreover, though studies have found anopheline species to be fairly homogenous over long ranges within a given environment, different ecoclimatological environments (e.g., dry savanna vs. humid coast) were found to drive selection, suggesting that the anopheline species may exhibit population structure across different environments (Lehmann et al. 1997). Given the evidence for population structure, and that the phylogenetic signals found in this data set are more complex than one expects to find assuming incomplete lineage sorting in panmictic ancestral populations alone, we deemed this a viable system in which to apply TASTI. In what follows, we conduct analyses of the autosomes for the *A. gambiae* complex, using a seventh species, *A. christyi*, as an outgroup.

A single individual from each of the six species was sequenced at high coverage. Wen, Yu, Hahn, et al. (2016) subset these high-coverage samples to generate independent genomic regions. Because TASTI also assumes independent loci, we used these alignments as input. After subsetting by Wen, Yu, Hahn, et al. (2016), an ample 2,791 independent alignments across the autosomes with an estimated mean length of 3.4 kb remained in this data set (figure S1 of Wen, Yu, Hahn, et al. [2016] shows a histogram of locus lengths). Following the procedures of both Fontaine et al. (2015) and Wen, Yu, Hahn, et al. (2016), we estimated maximum likelihood gene tree topologies under the GTR+Gamma model at each locus using RAxML (Stamatakis 2014). Supplementary figure S11, Supplementary Material online, displays the distributions of input rooted triples estimated in this manner, revealing asymmetries among some of the distributions of topologies. Following Wen, Yu, Hahn, et al. (2016), we also generated 100 bootstrap alignments at each locus and estimated the maximum likelihood topology for each bootstrap replicate. These bootstraps can be used to account for uncertainty in gene tree topology estimates, analogous to the Bayesian approach described by Yu et al. (2012). Briefly, let K be the number of loci, and for each locus k, let $g_{k1}, g_{k2}, \ldots, g_{kq}$ be the q different gene tree topologies estimated at that locus from the bootstrap alignments. Denote $\alpha_{k1}, \alpha_{k2}, \ldots, \alpha_{kq}$ as the proportion of times that gene tree topologies 1, $2, \ldots, q$ were, respectively, inferred from these alignments at locus k, such that $\sum_{i=1}^{q} \alpha_{ki} = 1$. Then, letting $\mathcal{G}$ be the set of all gene tree topologies computed across the bootstrap replicates across K loci, we define for each $g \in \mathcal{G}$ the sum of bootstrap probabilities associated with all loci whose topology is g as p_g. That is, for each $g \in \mathcal{G}$, we have $p_g = \sum_{k=1}^{K} \alpha_{kg}$ such that $\sum_{g \in \mathcal{G}} p_g = K$. Let $\mathbf{1}_{\{\mathcal{T} \text{ display } \mathcal{T}'\}}$ be an indicator random variable that takes the value 1 if tree topology $\mathcal{T}$ displays tree topology $\mathcal{T}'$, and 0 otherwise. We thus modify the likelihood of a species tree in equation (17) to be

$$L(\mathbf{\Omega}; \mathbf{y}) = \mathbb{P}[G = ((ab)c) | \mathbf{\Omega}]^{d_{((ab)c)}} \times \mathbb{P}[G = ((bc)a) | \mathbf{\Omega}]^{d_{((bc)a)}} \times \mathbb{P}[G = ((ac)b) | \mathbf{\Omega}]^{d_{((ac)b)}}, \tag{19}$$

where

$$\mathbf{y} = (d_{((ab)c)}, d_{((ac)b)}, d_{((bc)a)}) \tag{20}$$

and

$$d_{((xy)z)} = \sum_{g \in \mathcal{G}} p_g \mathbf{1}_{\{g \text{ displays } ((xy)z)\}} \tag{21}$$

denotes the sum of probabilities of gene trees displaying the rooted triple $((xy)z)$, such that $d_{((xy)z)} + d_{((xz)y)} + d_{((yz)x)} = K$.

To conduct our analyses, we first needed to obtain an estimate for the diversity parameter $\theta = 4N\mu$. Fontaine et al.

(2015) sequenced several individuals from each of the six species under study in the complex at lower coverage. We computed the mean pairwise sequence difference $\widehat{\pi}$ (Tajima 1983) within populations after filtering out indels using VCFtools (Danecek et al. 2011). Recalling that $\mathbb{E}[\widehat{\pi}] = \theta$ (Tajima 1983), we proceeded by using $\widehat{\pi}$ as an estimate of θ. Across the genome, however, the estimate of θ within each species in the complex varied by over an order of magnitude (from $\sim 1.4 \times 10^{-4}$ in *Anopheles melas* to $\sim 1.4 \times 10^{-3}$ in *A. arabiensis*). To approximate the ancestral estimates of θ, for any triple analyzed, we selected the maximum of the three corresponding estimates of θ as the value to pass to TASTI. The maximum was chosen to be robust to the bottleneck Fontaine et al. (2015) suspected to have occurred in the recent ancestry of some species in the complex.

Estimates of the effective population size of anopheline species are highly variable in the literature (Taylor et al. 1993; Michel et al. 2006; Hodges et al. 2013). We therefore considered supplying TASTI effective population sizes across different orders of magnitude. By fitting a model with $N \in \{10^3, 10^4, 10^5\}$, we found that our results were robust to effective population size. This is reasonable, since while the choice of θ directly affects the coalescent rate in the evolutionary process, the effective population size in essence only sets an upper bound on the migration rate. Analyses presented here assume an effective population size of $N = 10^4$ across all populations.

We applied TASTI to the original data set as well as 100 replicate data sets created by bootstrapping the original alignments. Parameter estimates with 95% confidence intervals for each estimated triple are shown in supplementary table S6, Supplementary Material online. Because $\widehat{M}$ lies on a continuum from complete structure to no structure, we used a testing procedure based on a parametric bootstrap to identify which triplets were estimated to be structured. The details of the testing procedure are in Algorithm 1, and the *P*-values indicating support for ancestral structure over the unstructured case are in supplementary table S6, Supplementary Material online. Note that even though we simulate bootstrap replicate gene tree sets under the best-fit structured population model from TASTI, the two least frequent gene tree topologies are set to have the same expected counts to mimic the symmetry expected under the null hypothesis of no structure. We applied this procedure to each set of rooted triples setting the number of bootstrap samples to $B = 100$.

Our results are concordant with earlier findings, indicating ancestral structure is frequently found among species belonging to the *A. gambiae* complex. We identified four triplets that are exceptions to this rule based on the testing procedure in Algorithm 1 with $\alpha = 0.01$: ((CG)L), ((CG)R), ((LR)C), and ((LR)G). These results support a claim that the pairs *A. coluzzii* and *A. gambiae*, and *A. melus* and *A. merus*, derive from panmictic ancestral populations. The full six-species estimated species tree topology with branch confidence determined by bootstraps is presented in figure 7. Our analysis returns the nearly fully resolved species tree ((*A. coluzzii*, *A. gambiae*), (*A. arabiensis*, (*A. melus*, *A. quadriannulatus*), *A. merus*)) with strong bootstrap support for the clades.

Algorithm 1 Testing $\mathcal{H}_0$:= panmictic populations versus $\mathcal{H}_1$:= nonpanmictic populations among taxa *x*, *y*, and *z* using a parametric bootstrap

```
Define $B$:= # of parametric bootstrap samples
for $b$ in 1 to $B$ do
    Simulate 2,791 gene tree topologies based on $\widehat{M}$ and $\widehat{\tau}_2 - \widehat{\tau}_1$ obtained from analysis of bootstrapped alignments (supplementary table S6, Supplementary Material online)
    Define observed counts $O^b = (n_{xy}, n_{xz}, n_{yz})$
    Define $i$:= argmax$\{O^b\}$
    Define expected counts $E^b = (\max\{O^b\}, \mathrm{mean}\{O^b_{-i}\}, \mathrm{mean}\{O^b_{-i}\})$
    Compute $P^b$ according to multinomial goodness-of-fit test based on the observed and expected counts using a $\chi^2$ distribution with 1 degree of freedom
Define $P_{xyz}$:= $P^1, \ldots, P^B$ combined into one $P$-value using Simes' procedure (Simes 1986)
Reject $\mathcal{H}_0$ when $P_{xyz} \leq \alpha$
```

In line with previous work, we estimated a well-supported clade between species *A. coluzzii* and *A. gambiae*. Though our analysis and the analyses of both Fontaine et al. (2015) and Wen, Yu, Hahn, et al. (2016) do not fully agree on the placement of the remaining four species in the phylogeny, because all three analyses assume a different model, the final results need not agree. It is notable, though, that the phylogeny inferred by TASTI in fact most resembles the phylogeny deduced from the X chromosome by Fontaine et al. (2015). The X chromosome of the anopheline species does not harbor signals of introgression (Fontaine et al. 2015), so the species tree constructed from those gene trees may be more reliable in that the underlying evolutionary process is simpler to tease out. That TASTI infers a phylogeny from the autosomal data which is similar to that of the X chromosome is a reassuring result; however, identifying the correct model to assume is a nontrivial task which deserves thorough investigation in the future.

Discussion

In this article, we developed TASTI, an algorithm that can be used to infer species trees from gene trees when a group of taxa's ancestral populations are structured, which offers the possibility to operate on larger numbers of taxa with a supertree approach. TASTI improves upon previous approaches (e.g., GLASS, STEM, and Maximum

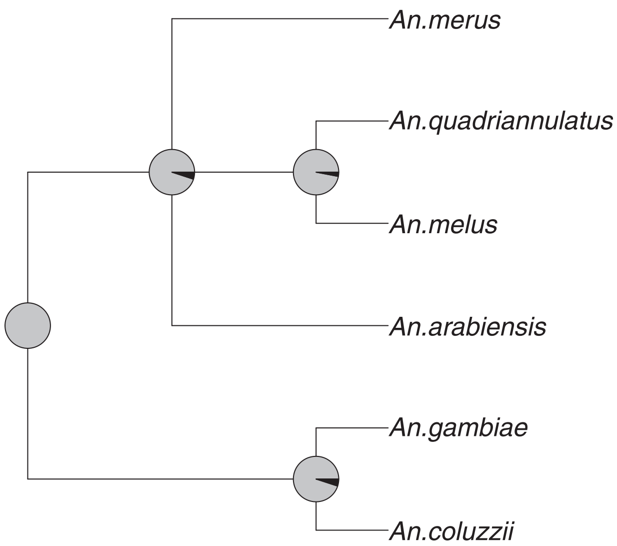

Fig. 7.—Species tree topology for the *Anopheles gambiae* complex estimated by applying the modified mincut supertree algorithm to the set of all distinct three-species tree topologies inferred by TASTI, with bootstrap branch support obtained from 100 bootstrap replicates. The proportion of gray in each pie chart corresponds to the degree of support for each clade. That is, a fully gray chart implies 100% bootstrap support for a given clade.

Tree) that accurately infer trees under this scenario (provided input trees are known with certainty) because it exhibits robustness to empirical data.

We recognize TASTI makes some simplifying assumptions about ancestral structure. Although the assumed persistent population structure may not be entirely realistic, it is plausible for this model to approximate signals of ancestral population structure that arise from organisms experiencing periodic separation based on environmental changes such as glaciations and sea level fluctuations (Toms et al. 2014) or limited offspring dispersal (Habets et al. 2006, 2007). Moreover, the assumption of no more than two subpopulations may for some species be insufficient. Additionally, even with only two subpopulations, one could forseeably introduce more complex species tree orientations by considering other ways in which an ancestry could be structured. However, for this model, such intricacies would come at the cost of total unidentifiability of the species tree model. This issue for phylogenetic inference has been discussed at length and is not unique to TASTI (Chang 1996; Evans and Warnow 2004; Ho and Ané 2014; Xu and Yang 2016).

A substantial amount of recent research has focused on advancing models for estimating reticulate evolutionary histories caused by events such as hybridization and continuous gene flow (Huson et al. 2005; Baum 2007; Meng and Kubatko 2009; Gerard et al. 2011; Yu et al. 2014; Solís-Lemus and Ané 2016; Tian and Kubatko 2016; Wen and Nakhleh 2017; Hey et al. 2018; Long and Kubatko 2018). To contrast, our extensive simulations support the idea that asymmetries in gene tree topology distributions not observed under the standard multispecies coalescent, which are frequently attributed to interspecies gene flow, also arise under ancestral population structure (Slatkin and Pollack 2008). However, such asymmetries in gene tree topology distributions are likely to be generated only under settings where the degree M of gene flow relative to the time τ between species divergences is small. Consider the internal branch of our structured population three-taxon model with topology ((AB)C). The waiting time to the first migration event (either from the first or the second subpopulation) is exponentially distributed with rate $2m$, where m is the per-generation migration rate. Following DeGiorgio and Rosenberg (2016), the probability of no migration event on the internal branch of length t generations (i.e., neither lineage migrates) is $\exp(-2mt) = \exp(-2M\tau)$, where $M = 4Nm/\theta$ and $\tau = t\theta/(2N)$ are scaled in mutation units. Moreover, the probability that the first event above the root is a coalescence event that would lead to gene tree topology $((bc)a)$ is $\exp(-2mt)/(6Nm+1) = \exp(-2M\tau)/(3M\theta/2+1)$, which is arbitrarily close to one for M small enough with fixed τ. This result indicates that asymmetries in gene tree distributions will manifest in scenarios when $M\tau$ is small—that is, when gene flow between subpopulations is infrequent on the time scale separating species splits.

Though population structure is often explored in analyses of evolutionary histories (e.g., F-statistics such as F_{ST} [Wright 1949; Weir and Cockerham 1984] and software such as STRUCTURE [Pritchard et al. 2000] for exploring population structure enjoy widespread use), model-based approaches incorporating such structure within species inference frameworks have been lacking. Our analyses of the *A. gambiae* complex data further demonstrate, when comparing to the work of Wen, Yu, Hahn, et al. (2016) and Fontaine et al. (2015), that similar phylogenies can be inferred when assuming either a reticulate history or ancestral population structure. Indeed, the identifiability of such histories given gene tree topologies alone is severely limited or impossible. An investigation via comprehensive simulation studies of how distributions of gene tree branch lengths differ under various evolutionary assumptions such as hybridization and population structure may provide key insights into selection of an appropriate evolutionary model.

A promising alternative, especially in the absence of external information, would reformulate TASTI in terms of a hypothesis test. That is, it may be meaningful to test for significant confidence in ancestral structure versus unstructured populations using a bootstrapping approach. Moreover, a different test could assess the likelihood of ancestral structure versus hybridization, though addressing such a question may be challenging based on topology data alone. Further work remains in examining how phylogenetic signals differ under these various modeling assumptions, such that an appropriate model may be determined on a case-by-case basis.

If possible, external information or expert opinion can be used to aid in model performance. In the case of identifiability, external knowledge can be supplied to restrict the domain of certain free parameters, namely the divergence times, such that the model becomes identifiable (Huelsenbeck et al. 2008). Models equally likely to produce a given distribution of gene tree topologies can further be evaluated by more closely examining branch lengths or molecular sequences (Yu and Nakhleh 2015). External information in the form of Bayesian prior distributions can boost the performance of gene tree inference as well (Huelsenbeck and Ronquist 2001; Huelsenbeck et al. 2004). The more accurate the input data, the closer TASTI gets to achieving optimal performance. This is especially crucial when using data on gene trees with branch lengths. Simultaneous estimation of genes trees and the species tree from multilocus data in a Bayesian hierarchical framework has also proven effective and popular (Liu and Pearl 2007; Liu 2008; Heled and Drummond 2010; Wen, Yu, and Nakhleh 2016; Zhang et al. 2017), and an extension of TASTI to such a framework may be powerful.

Some avenues are available to reduce computation time and improve performance in analyses with larger numbers of species. Maximum pseudolikelihood methods based on groups of rooted triples have been proposed both for the standard multispecies coalescent with incomplete lineage sorting (Liu, Yu, and Pearl 2010) and with hybridization (Yu and Nakhleh 2015; Solís-Lemus and Ané 2016). These approaches have been shown to reduce computational burden, an important feature given that, even with greatly diminishing the computational task using a supertree approach, TASTI's computation time grows with the number of taxa at a rate of $\binom{n}{3}$. Thus, a pseudolikelihood method for taxa with ancestral population structure would be a useful extension to our work. Finally, although we find favorable performance of the modified mincut supertree algorithm of Page (2002), the effect of alternative supertree construction algorithms (Wilkinson et al. 2005) could nevertheless be explored.

Data Availability

Independent alignments of high-depth samples from the *Anopheles gambiae* complex, first generated by Wen, Yu, Hahn, et al. (2016), are available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.tn47c. The VCF files for low-depth samples from the *A. gambiae* complex, detailed by Fontaine et al. (2015), are also available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.f4114. An R package, called TASTI, containing the functions used to implement the core analyses presented here is available for download from https://github.com/hillarykoch/TASTI.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by National Science Foundation (Grant Nos. DEB-1753489, DEB-1949268, and BCS-2001063), National Institutes of Health (Grant No. R35-GM128590), the Alfred P. Sloan Foundation, a National Institute of General Medical Sciences-funded training grant on Computation, Bioinformatics and Statistics (Predoctoral Training Program T32GM102057), and by a National Human Genome Research Institute pre-doctoral fellowship (1F31HG010574-01). Portions of this research were conducted with Advanced CyberInfrastructure computational resources provided by the Institute for CyberScience at Pennsylvania State University.

Literature Cited

Allman ES, Degnan JH, Rhodes JA. 2011. Determining species tree topologies from clade probabilities under the coalescent. J Theor Biol. 289:96–106.

Baum DA. 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. Taxon 56(2):417–426.

Burgess R, Yang Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol Biol Evol. 25(9):1979–1994.

Casella G, Berger RL. 2002. Statistical inference. Vol. 2. Pacific Grove (CA): Duxbury.

Chang JT. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math Biosci. 137(1):51–73.

Cranston KA, Hurwitz B, Ware D, Stein L, Wing RA. 2009. Species trees from highly incongruent gene trees in rice. Syst Biol. 583:489–600.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics. 27(15):2156–2158.

Davidson G. 1962. *Anopheles gambiae* complex. Nature 196(4857):907.

DeGiorgio M, et al. 2014. An empirical evaluation of two-stage species tree inference strategies using a multilocus dataset from North American pines. BMC Evol Biol. 14(1):67.

DeGiorgio M, Degnan JH. 2010. Fast and consistent estimation of species trees using supermatrix rooted triples. Mol Biol Evol. 27(3):552–569.

DeGiorgio M, Degnan JH. 2014. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Syst Biol. 64:66–82.

DeGiorgio M, Rosenberg NA. 2016. Consistency and inconsistency of consensus methods for estimating species trees from gene trees. Theor Popul Biol. 110:12–24.

Degnan JS, DeGiorgio M, Bryant D, Rosenberg NA. 2009. Properties of consensus methods for inferring species trees from gene trees. Syst Biol. 58(1):35–54.

Drummond AJ, Rambaut A. 2007. Beast: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7(1):214.

Escobar D, Zea S, Sánchez JA. 2012. Phylogenetic relationships among the Caribbean members of the *Cliona viridis* complex (Porifera, Demospongiae, Hadromerida) using nuclear and mitochondrial DNA sequences. Mol Phylogenet Evol. 64(2):271–284.

Evans SN, Warnow T. 2004. Unidentifiable divergence times in rates-across-sites models. IEEE/ACM Trans Comput Biol Bioinform. 1(3):130–134.

Felsenstein J. 1989. Phylip—phylogeny inference package. Cladistics 5:164–166.

Felsenstein J. 2004. Inferring phylogenies. Vol. 2. Sunderland (MA): Sinauer Associates.

Fontaine MC, et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347(6217):1258524.

Garrigan D, Mobasher Z, Kingan SB, Wilder JA, Hammer MF. 2005. Deep haplotype divergence and long-range linkage disequilibrium at xp21.1 provide evidence that humans descend from a structured ancestral population. Genetics 170(4):1849–1856.

Gerard D, Gibbs HL, Kubatko L. 2011. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. BMC Evol Biol. 11(1):291.

Habets MG, Czaran T, Hoekstra RF, De Visser JAG. 2007. Spatial structure inhibits the rate of invasion of beneficial mutations in asexual populations. Proc R Soc B. 274(1622):2139–2143.

Habets MG, Rozen DE, Hoekstra RF, de Visser JAG. 2006. The effect of population structure on the adaptive radiation of microbial populations evolving in spatially structured environments. Ecol Lett. 9(9):1041–1048.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 22(2):160–174.

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

Helmkamp LJ, Jewett EM, Rosenberg NA. 2012. Improvements to a class of distance matrix methods for inferring species trees from gene trees. J Comput Biol. 19(6):632–649.

Hey J, et al. 2018. Phylogeny estimation by integration over isolation with migration models. Mol Biol Evol. 35(11):2805–2818.

Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci U S A. 104(8):2785–2790.

Ho LST, Ané C. 2014. Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. Methods Ecol Evol. 5(11):1133–1146.

Hobolth A, Norvang Andersen L, Mailund T. 2011. On computing the coalescence time density in an isolation-with-migration model with few samples. Genetics 187(4):1241–1243.

Hodges TK, et al. 2013. Large fluctuations in the effective population size of the malaria mosquito *Anopheles gambiae* s.s. during vector control cycle. Evol Appl. 6(8):1171–1183.

Hudson RR. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37(1):203–217.

Hudson RR. 2002. Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics 18(2):337–338.

Huelsenbeck JP, Joyce P, Lakner C, Ronquist F. 2008. Bayesian analysis of amino acid substitution models. Philos Trans R Soc B. 363(1512):3941–3953.

Huelsenbeck JP, Larget B, Alfaro ME. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol Biol Evol. 21(6):1123–1133.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17(8):754–755.

Huson DH, Klöpper T, Lockhart PJ, Steel MA. 2005. Reconstruction of reticulate networks from gene trees. In: Annual International Conference on Research in Computational Molecular Biology; 2005 May 14; Berlin, Heidelberg: Springer. p. 233–249.

Jewett EM, Rosenberg NA. 2012. iGLASS: an improvement to the GLASS method for estimating species trees from gene trees. J Comput Biol. 19(3):293–315.

Johnson BR, et al. 2013. Phylogenomics resolves evolutionary relationships among ants, bees, and wasps. Curr Biol. 23(20):2058–2062.

Kingman JFC. 1982. The coalescent. Stoch Proc Appl. 13(3):235–248.

Kubatko LS, Carstens BC, Knowles LL. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25(7):971–973.

Leaché AD, Harris RB, Rannala B, Yang Z. 2014. The influence of gene flow on species tree estimation: a simulation study. Syst Biol. 63(1):17–30.

Leaché AD, Rannala B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Syst Biol. 60(2):126–137.

Lehmann T, et al. 1997. Microgeographic structure of *Anopheles gambiae* in western Kenya based on mtDNA and microsatellite loci. Mol Ecol. 6(3):243–253.

Lehmann T, Hawley WA, Grebert H, Collins FH. 1998. The effective population size of *Anopheles gambiae* in Kenya: implications for population structure. Mol Biol Evol. 15(3):264–276.

Lin M, et al. 2014. Mitochondrial genome rearrangements in the scleractinia/corallimorpharia complex: implications for coral phylogeny. Genome Biol Evol. 6(5):1086–1095.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24(21):2542–2543.

Liu L, Pearl DK. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol. 56(3):504–514.

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

Liu L, Yu L, Pearl DK. 2010. Maximum tree: a consistent estimator of the species tree. J Math Biol. 60(1):95–106.

Long C, Kubatko LS. 2018. The effect of gene flow on coalescent-based species-tree inference. Syst Biol. 67(5):770–785.

Mailund T, et al. 2012. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genet. 8(12):e1003125.

Marcussen T, et al. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. New Phytol. 345(6194):1250092.

McGuire JA, et al. 2007. Mitochondrial introgression and incomplete lineage sorting through space and time: phylogenetics of crotaphytid lizards. Mol Phylogenet Evol. 61(12):2879–2897.

Meng C, Kubatko LS. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor Popul Biol. 75(1):35–45.

Michel AP, et al. 2006. Effective population size of *Anopheles funestus* chromosomal forms in Burkina Faso. Malar J. 5(1):115.

Mirarab S, et al. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30(17):i541–i548.

Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31(12):i44–i52.

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

Nakhleh L. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol Evol. 28(12):719–728.

Neafsey DE, et al. 2015. Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. Science 347(6217):1258522.

Page RDM. 2002. Modified mincut supertrees. In: Guigó R, Gusfield D, editors. Proceedings of the Second International Workshop on Algorithms in Bioinformatics (WABI, 2002 September 17). Vol. 2452 (Lecture Notes in Computer Science). Berlin (Germany). Berlin, Heidelberg: Springer. p. 537–551.

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

Pei J, Wu Y. 2017. STELLS2: fast and accurate coalescent-based maximum likelihood inference of species trees from gene tree topologies. Bioinformatics 33(12):1789–1797.

Peters RS, et al. 2017. Evolutionary history of the hymenoptera. Curr Biol. 27(7):1013–1018.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155(2):945–959.

Rambaut A, Grass NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics 13(3):235–238.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci. 53(1–2):131–147.

Ross SM. 2014. Introduction to probability models. Cambridge: Academic Press.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497(7449):327–331.

Semple C, Steel MA. 2003. Phylogenetics. Vol. 24. Oxford: Oxford University Press.

Shen X, et al. 2016. Reconstructing the backbone of the *Saccharomycotina* yeast phylogeny using genome-scale data. G3 6(12):3927–3939.

Simes RJ. 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73(3):751–754.

Slatkin M, Pollack JL. 2008. Subdivision in an ancestral species creates asymmetry in gene trees. Mol Biol Evol. 25(10):2241–2246.

Solís-Lemus C, Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet. 12(3):e1005896.

Solís-Lemus C, Yang M, Ané C. 2016. Inconsistency of species tree methods under gene flow. Syst Biol. 65(5):843–851.

Song S, Liu L, Edwards SV, Wu S. 2012. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS One 8:e54848.

Song S, et al. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci U S A. 109(37):14942–14947.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.

Steel MA. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. J Classif. 9(1):91–116.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105(2):437–460.

Taylor CE, Toure YT, Coluzzi M, Petrarca V. 1993. Effective population size and persistence of *Anopheles arabiensis* during the dry season in West Africa. Med Vet Entomol. 7(4):351–357.

Thalmann O, Fischer A, Lankester F, Paabo S, Vigilant L. 2006. The complex evolutionary histories of gorillas: insights from genomic data. Mol Biol Evol. 24(1):146–158.

Than C, Ruths D, Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary histories. BMC Bioinformatics 9(1):322.

Tian Y, Kubatko LS. 2016. Distribution of coalescent histories under the coalescent model with gene flow. Mol Phylogenet Evol. 105:177–192.

Toms JA, Compton JS, Smale M, von der Heyden S. 2014. Variation in palaeo-shorelines explains contemporary population genetic patterns of rocky shore species. Biol Lett. 10(6):20140330.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38(6):1358–1370.

Wen D, Nakhleh L. 2017. Co-estimating reticulate phylogenies and gene trees from multi-locus sequence data. Syst Biol. 67(3):439–457.

Wen D, Yu Y, Hahn MW, Nakhleh L. 2016. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Mol Ecol. 25(11):2361–2372.

Wen D, Yu Y, Nakhleh L. 2016. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. PLoS Genet. 12(5):e1006006.

White MA, Ané C, Dewey CN, Larget BR, Payseur BA. 2009. Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genet. 5(11):e1000729.

Wilkinson M, et al. 2005. The shape of supertrees to come: tree shape related properties of fourteen supertree methods. Syst Biol. 54(3):419–431.

Wright S. 1949. The genetical structure of populations. Ann Eugen. 15(4):323–354.

Wu Y. 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution 66(3):763–775.

Xu B, Yang Z. 2016. Challenges in species tree estimation under the multispecies coalescent model. Genetics 204(4):1353–1368.

Yang Y, et al. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. Mol Biol Evol. 32(8):2001–2014.

Yu Y, Degnan JH, Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet. 8(4):e1002660.

Yu Y, Dong J, Liu KJ, Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proc Natl Acad Sci U S A. 111(46):16448–16453.

Yu Y, Nakhleh L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics. 16(S10):S10.

Yu Y, Than C, Degnan JH, Nakhleh L. 2011. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst Biol. 60(2):138–149.

Zhang C, Ogilvie HA, Drummond AJ, Stadler T. 2017. Bayesian inference of species networks from multilocus sequence data. Mol Biol Evol. 35(2):504–517.

Associate editor: David Enard