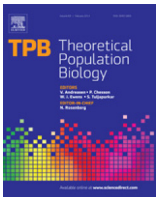
Theoretical Population Biology 110 (2016) 12–24

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb



Consistency and inconsistency of consensus methods for inferring species trees from gene trees in the presence of ancestral population structure

Michael DeGiorgio [a,*], Noah A. Rosenberg [b]

[a] Department of Biology, Pennsylvania State University, 502 Wartik Laboratory, University Park, PA 16802, USA

[b] Department of Biology, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 28 January 2014
Available online 13 April 2016

Keywords:
Coalescent
Consensus methods
Incomplete lineage sorting
Phylogenetics
Statistical consistency

ABSTRACT

In the last few years, several statistically consistent consensus methods for species tree inference have been devised that are robust to the gene tree discordance caused by incomplete lineage sorting in unstructured ancestral populations. One source of gene tree discordance that has only recently been identified as a potential obstacle for phylogenetic inference is ancestral population structure. In this article, we describe a general model of ancestral population structure, and by relying on a single carefully constructed example scenario, we show that the consensus methods Democratic Vote, STEAC, STAR, R* Consensus, Rooted Triple Consensus, Minimize Deep Coalescences, and Majority-Rule Consensus are statistically inconsistent under the model. We find that among the consensus methods evaluated, the only method that is statistically consistent in the presence of ancestral population structure is GLASS/Maximum Tree. We use simulations to evaluate the behavior of the various consensus methods in a model with ancestral population structure, showing that as the number of gene trees increases, estimates on the basis of GLASS/Maximum Tree approach the true species tree topology irrespective of the level of population structure, whereas estimates based on the remaining methods only approach the true species tree topology if the level of structure is low. However, through simulations using species trees both with and without ancestral population structure, we show that GLASS/Maximum Tree performs unusually poorly on gene trees inferred from alignments with little information. This practical limitation of GLASS/Maximum Tree together with the inconsistency of other methods prompts the need for both further testing of additional existing methods and development of novel methods under conditions that incorporate ancestral population structure.

1. Introduction

Recently, much attention has been given to the development of methods that consistently infer the correct species tree from the discordant gene trees produced under incomplete lineage sorting—the failure of lineages from two different species to coalesce in the population immediately ancestral to the divergence of the two species (Degnan and Rosenberg, 2009). Consensus approaches, each of which takes a set of gene trees as input and returns a species tree estimate according to a specific rule (Bryant, 2003), have provided one important source of methods for species tree inference in this context.

A consensus method $\widehat{C}$ is a statistically consistent estimator of a species tree topology under some model if for each species tree σ, $\widehat{C}$ applied to a set of gene trees randomly generated under the model, assuming that the species tree is σ, converges in probability to the topology of σ as the number of gene trees approaches ∞. Statistical consistency is a desirable property because it is reasonable to expect that as more data are gathered, evidence should accumulate in support of the true value of the parameter being estimated.

Degnan and Rosenberg (2006) showed that when gene trees are distributed according to the multispecies coalescent model for the evolution of gene lineages conditional on a species tree, an extreme case of incomplete lineage sorting can arise in which the most likely gene tree topology does not match the species tree topology. This inconsistency implies that species tree estimation methods must use information other than the most frequently occurring gene tree topology in order to accurately infer the

[*] Corresponding author.
E-mail address: mxd60@psu.edu (M. DeGiorgio).

http://dx.doi.org/10.1016/j.tpb.2016.02.002
0040-5809/© 2016 Elsevier Inc. All rights reserved.

Table 1
Notation.

Notation	Definition
$\mathbf{D}$	$(n-1)$-dimensional vector of the numbers of demes in the $n-1$ ancestral populations
$\mathbf{N}$	$(n-1)$-dimensional vector with vector-valued elements for the deme sizes in each of the $n-1$ ancestral populations
$\mathbf{M}$	$(n-1)$-dimensional vector with matrix-valued elements for the backward migration matrices in each of the $n-1$ ancestral populations
$\boldsymbol{\Psi}$	Matrix that describes how demes connect across species boundaries
$\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$	Ancestral population structure model with parameters σ, $\mathbf{D}$, $\mathbf{N}$, $\mathbf{M}$, and $\boldsymbol{\Psi}$
$\mathbb{P}[E\,;\,\mathcal{S}]$	Probability of event E under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$
λ_{A}	Subtree of species tree σ that contains species A and that descends from the divergence of species A and B
λ_{B}	Subtree of species tree σ that contains species B and that descends from the divergence of species A and B
λ_{C}	Subtree of species tree σ that contains species C and that descends from the divergence of species (AB) and C
$\Gamma_{\mathrm{A}}, \Gamma_{\mathrm{B}}, \Gamma_{\mathrm{C}}$	Sets of taxa at the leaves of subtrees λ_{A}, λ_{B}, and λ_{C}, respectively
$\mathcal{L}$	Set of taxa
$\mathcal{T} \mid \mathcal{L}$	Tree displayed by phylogenetic tree $\mathcal{T}$ restricted to the set of taxa $\mathcal{L}$
$\mathrm{top}(\mathcal{T})$	Topology of phylogenetic tree $\mathcal{T}$
$p_{\mathcal{S}}(\mathrm{X}, \mathrm{Y})$	Probability that a lineage sampled from species X and a lineage sampled from species Y are in the same deme at the speciation time of X and Y under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$
$P_{\mathcal{S}}[\mathcal{T}]$	Probability of gene tree topology $\mathcal{T}$ under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$
$\widehat{P}[\mathcal{T}]$	Sample proportion of topology $\mathcal{T}$ in a set of gene trees
T_{XY}^{ℓ}	Random coalescence time at locus ℓ for a lineage sampled from species X and a lineage sampled from species Y
$\mathbb{E}_{\mathcal{S}}[T_{\mathrm{XY}}^{\ell}]$	Expected coalescence time under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$ for a lineage sampled from species X and a lineage sampled from species Y at locus ℓ
$\overline{T}_{\mathrm{XY}}$	Mean coalescence time across all sampled gene trees between one lineage sampled from species X and one lineage sampled from species Y
R_{XY}^{ℓ}	Rank of the coalescent event at locus ℓ for a lineage sampled from species X and a lineage sampled from species Y
$\mathbb{E}_{\mathcal{S}}[R_{\mathrm{XY}}^{\ell}]$	Expected rank under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$ of the coalescent event for a lineage sampled from species X and a lineage sampled from species Y at locus ℓ
$\overline{R}_{\mathrm{XY}}$	Mean rank of coalescent events across all sampled gene trees between one lineage sampled from species X and one lineage sampled from species Y
$t_{\mathrm{XY}}^{\min}$	Minimum coalescence time across all sampled gene trees between one lineage sampled from species X and one lineage sampled from species Y
$\mathrm{xl}(\mathrm{top}(\sigma), \mathcal{T})$	Number of extra lineages contributed by the topology of fixed species tree σ for a fixed gene tree topology $\mathcal{T}$
$\mathrm{xl}(\mathrm{top}(\sigma))$	Number of extra lineages contributed by the topology of fixed species tree σ for a fixed set of gene trees

species tree topology. Indeed, many consensus methods relying on other principles provide statistically consistent estimators of the species tree topology under the multispecies coalescent model. This collection of methods includes STEAC (Liu et al., 2009), STAR (Liu et al., 2009), R* Consensus (Degnan et al., 2009), GLASS (Mossel and Roch, 2010), and Maximum Tree (Liu et al., 2010), as well as extensions of some of these methods that preserve the consistency property (Helmkamp et al., 2012; Jewett and Rosenberg, 2012; Allman et al., 2013).

In its simplest form, the multispecies coalescent model assumes that each modern species and each ancestral species have a constant population size, each pair of lineages within a given ancestral species has an equal chance of coalescing, and each species is an unstructured population. Because the multispecies coalescent assumes that random mating occurs within species, when ancestral species are structured, as has been argued for various species (e.g., Garrigan et al., 2005; Thalmann et al., 2007; White et al., 2009), it is unclear whether methods that are consistent under the multispecies coalescent continue to be consistent.

The difficulty of species tree estimation in the presence of ancestral population structure lies in the way that population structure alters the probability distribution of gene trees given a species tree compared to the unstructured case. Using a three-taxon example, Slatkin and Pollack (2008) showed that with ancestral population structure, the probability distribution of gene tree topologies can have a certain asymmetry, and the most likely three-taxon gene tree topology need not match the species tree topology. These consequences of the multispecies coalescent with ancestral structure do not occur in the standard multispecies coalescent.

Here, we describe an extension of the ancestral population structure model considered by Slatkin and Pollack (2008). Using our extended model, we evaluate the consistency of several consensus methods, employing a single example scenario to show that many methods are inconsistent. We show that each of the inconsistent methods is in fact "misleading" in the sense that for a certain fixed species tree σ and a particular set of parameters, the probability that the consensus tree contains a clade not present on σ approaches 1 as the number of loci approaches ∞. To evaluate the speed at which methods converge to or diverge from the correct bifurcating species tree topology, we perform simulations of our model. As predicted by our theoretical results, the only method that does not strongly support incorrect species tree topologies is GLASS/Maximum Tree. However, in accord with past simulations using model species trees (Liu et al., 2009; Leaché and Rannala, 2011; Wu, 2012; DeGiorgio and Degnan, 2014), we show that GLASS/Maximum Tree performs poorly when an absence of substitutions causes little information to exist in sequence alignments. We conclude with a discussion of the implications of the results for understanding evolutionary relationships.

2. Model

We use the notation in Table 1. Suppose time is measured in generations, and that generation time is constant throughout the tree. Consider an ultrametric n-taxon bifurcating species tree σ with $n \geq 3$ taxa (i.e., each leaf has an identical sum of branch lengths to the root). Then we can always find a set of species A, B, and C on σ with relationship $((\mathrm{A}{:}\tau_3, \mathrm{B}{:}\tau_3){:}\tau_2 - \tau_3, \mathrm{C}{:}\tau_2)$, where $\tau_2 > \tau_3 > 0$.

Each internal branch along the species tree specifies an ancestral population. An n-taxon species tree contains $n - 1$ such populations, including the branch above the root. Label these populations of σ by recursively visiting the root, then the left subtree, and finally the right subtree (a pre-order traversal of σ). Each ancestral population is allowed to be structured; the population structure model is identical across L independent loci, so that each of L gene trees is a random variate conditional on the same species tree.

In ancestral population i, let $D^{(i)}$ be the number of demes, let $\mathbf{N}^{(i)}$ be the vector of population sizes for the $D^{(i)}$ demes, and let $\mathbf{M}^{(i)}$ be the backward migration matrix between demes (Fig. 1). Denote the ancestral population structure model by $\mathcal{S} = \mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$, where $\mathbf{D} = [D^{(1)}, D^{(2)}, \ldots, D^{(n-1)}]$, $\mathbf{N} = [\mathbf{N}^{(1)}, \mathbf{N}^{(2)}, \ldots, \mathbf{N}^{(n-1)}]$, $\mathbf{M} = [\mathbf{M}^{(1)}, \mathbf{M}^{(2)}, \ldots, \mathbf{M}^{(n-1)}]$, and $\boldsymbol{\Psi}$ is an $(n + \sum_{i=1}^{n-1} D^{(i)}) \times$

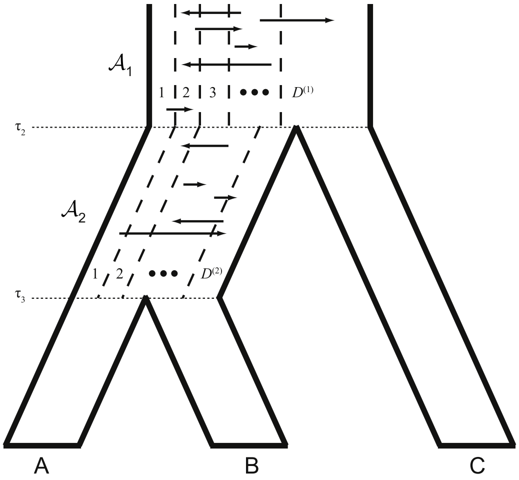


Fig. 1. Model for the relationship among species A, B, and C in a fixed species tree σ. Ancestral population $\mathcal{A}_1$ has $D^{(1)}$ demes and ancestral population $\mathcal{A}_2$ has $D^{(2)}$ demes. Migration occurs between the $D^{(1)}$ demes in ancestral population $\mathcal{A}_1$ and between the $D^{(2)}$ demes in ancestral population $\mathcal{A}_2$. At τ_2 and τ_3, going back in time, lineages merge into specific demes in ancestral populations $\mathcal{A}_1$ and $\mathcal{A}_2$ according to entries of the matrix $\mathbf{\Psi}$ (see Section 2).

$(n + \sum_{i=1}^{n-1} D^{(i)})$ matrix that describes how demes connect across species divergences. Each row (column) of $\mathbf{\Psi}$ corresponds to a distinct deme in an extant or ancestral population. The first n rows (columns) correspond to the n extant populations, the next $D^{(1)}$ rows (columns) to the $D^{(1)}$ demes in ancestral population 1, the next $D^{(2)}$ rows (columns) to the $D^{(2)}$ demes in ancestral population 2, and so on, until the last $D^{(n-1)}$ rows (columns) correspond to the $D^{(n-1)}$ demes in ancestral population $n-1$. Each extant population contains only a single deme because it is unstructured. The entry $\mathbf{\Psi}_{jk}$ provides the probability that a lineage merges into deme k from deme j in the moment at which, going back in time, deme j ends and deme k begins. If, going back in time, deme k does not directly receive lineages from deme j at this transition point, then $\mathbf{\Psi}_{jk} = 0$. By construction, each row of $\mathbf{\Psi}$ sums to 1. Therefore, $\mathbf{\Psi}$ provides probability distributions on the locations in the ancestral populations of σ into which lineages sampled from the taxa of σ can merge.

For each ancestral population i, $\mathbf{M}_{xy}^{(i)}$ is the per-generation probability of backward migration to deme $y^{(i)}$ for a lineage in deme $x^{(i)}$. Assume that in any ancestral population i, demes $x^{(i)}$ and $y^{(i)}$ communicate; that is, the migration rate from deme $x^{(i)}$ to deme $y^{(i)}$ is nonzero, or otherwise, for any pair of lineages there exists an indirect migration path through other demes from deme $x^{(i)}$ to deme $y^{(i)}$. This assumption encodes the idea of what we mean by a structured population rather than a series of separate populations; by ensuring that demes communicate in the ancestral population above the root, it guarantees that with probability 1, the coalescence process terminates. The relationship between species A, B, and C within the n-taxon species tree σ, and the ancestral population structure model, are illustrated in Fig. 1.

We are interested in the probabilities $\mathbb{P}[E\,;\,\mathcal{S}]$ of events E under model $\mathcal{S}$. Such probabilities are possible to compute by connecting models of individual populations along branches of species tree σ with rules given by model $\mathcal{S}$ about what happens to lineages at species divergence times. Note that in our calculations, we sample only one lineage from each extant population; it is then convenient to assume that the extant populations are unstructured. With one lineage sampled per population, coalescences take place only in ancestral populations, and the demographic model in extant populations does not affect patterns of coalescence.

3. Example scenario

We introduce a specific scenario and use it to prove that in part of the parameter space of our model, Democratic Vote (Degnan and Rosenberg, 2006, 2009; Rosenberg, 2013), STAR (Liu et al., 2009), STEAC (Liu et al., 2009), R* Consensus (Bryant, 2003; Degnan et al., 2009), Rooted Triple Consensus (Ewing et al., 2008), Minimize Deep Coalescences (MDC; Maddison, 1997; Maddison and Knowles, 2006; Than and Nakhleh, 2009; Nakhleh, 2013), and Majority-Rule Consensus (Degnan et al., 2009) are misleading in that the probability that a consensus tree contains a clade not on the species tree approaches 1 as the number of loci goes to ∞. Note that we only explore a subset of the available methods for inferring species trees from gene trees, and that the asymptotic properties of other methods could be investigated using similar approaches.

Consider a sample of n individuals, one from each species within an n-taxon species tree σ. Fig. 2(A) displays a set of three species A, B, and C that have the topological relationship ((AB)C) within σ. Certain internal branches are made long so that σ resembles a three-taxon species tree, in the sense that coalescences of lineages from the $n-3$ taxa other than A, B, and C with lineages from A, B, and C are likely to occur on these long internal branches (Fig. 2(B)).

Let λ_{A} be the subtree of σ that contains species A and that descends from the divergence of species A and B, let λ_{B} be the subtree of σ that contains species B and that descends from this same divergence, and let λ_{C} be the subtree of σ that contains species C and that descends from the divergence of species (AB) and C. Further, let $\mathit{\Gamma}_{\mathrm{A}}$, $\mathit{\Gamma}_{\mathrm{B}}$, and $\mathit{\Gamma}_{\mathrm{C}}$ denote the sets of taxa at the leaves of subtrees $\lambda_{\mathrm{A}}, \lambda_{\mathrm{B}}$, and λ_{C}, respectively. By definition, $\mathit{\Gamma}_{\mathrm{A}} \cap \mathit{\Gamma}_{\mathrm{B}} = \emptyset$, $\mathit{\Gamma}_{\mathrm{A}} \cap \mathit{\Gamma}_{\mathrm{C}} = \emptyset$, $\mathit{\Gamma}_{\mathrm{B}} \cap \mathit{\Gamma}_{\mathrm{C}} = \emptyset$, and $\mathit{\Gamma}_{\mathrm{A}} \cup \mathit{\Gamma}_{\mathrm{B}} \cup \mathit{\Gamma}_{\mathrm{C}}$ is the set of all taxa on species tree σ. Given a set of taxa $\mathcal{L}$, we denote by $\mathcal{T}|\mathcal{L}$ the tree displayed by tree $\mathcal{T}$ restricted to $\mathcal{L}$. We denote the topology of tree $\mathcal{T}$ by $\mathrm{top}(\mathcal{T})$. To show that a consensus method is misleading, it suffices to find a set of branch lengths on a fixed species tree σ such that as the number of loci approaches ∞, the probability approaches 1 that the inferred species tree contains a clade not on σ.

In our example scenario, we suppose that certain internal branches are long enough that for a fixed subset of taxa $\mathcal{L}$, where $\mathcal{L}$ is either $\mathit{\Gamma}_{\mathrm{A}}$, $\mathit{\Gamma}_{\mathrm{B}}$, or $\mathit{\Gamma}_{\mathrm{C}}$, fixed species tree σ, and random gene tree $\mathcal{T}$, the probability $\mathbb{P}[\mathrm{top}(\mathcal{T}|\mathcal{L}) = \mathrm{top}(\sigma|\mathcal{L})\,;\,\mathcal{S}]$ that $\mathcal{T}$ and σ have the same topology when restricted to the set of taxa $\mathcal{L}$ can be made arbitrarily close to 1. Formally, for fixed arbitrarily small $\delta > 0$, we make certain internal branches long enough that $1 - \delta < \mathbb{P}[\mathrm{top}(\mathcal{T}|\mathcal{L}) = \mathrm{top}(\sigma|\mathcal{L})\,;\,\mathcal{S}] < 1$. Therefore, because a random gene tree can be made to display the subtrees λ_{A}, λ_{B}, and λ_{C} with probability near 1, to prove that a consensus method is a misleading estimator of $\mathrm{top}(\sigma)$, it suffices to show that the species tree estimate on the basis of the method is not $((\lambda_{\mathrm{A}}\lambda_{\mathrm{B}})\lambda_{\mathrm{C}})$ in the limit as the number of gene trees goes to ∞.

Two ancestral populations are of interest, $\mathcal{A}_2$, directly ancestral to the divergence of species A and B at time τ_3, and $\mathcal{A}_1$, directly ancestral to the divergence of species (AB) and C at time τ_2. Ancestral populations $\mathcal{A}_1$ and $\mathcal{A}_2$ each contain $D^{(1)} = D^{(2)} = D \geq 2$ demes, each of size N diploid individuals, that exchange migrants according to migration matrices $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$. Note that we use D to indicate the number of demes in each ancestral population, whereas $\mathbf{D}$ denotes the $(n-1)$-dimensional vector, each of whose elements gives the number of demes in the associated ancestral population. For simplicity, both for $i = 1$ and $i = 2$, we assume a symmetric island migration model in which $\mathbf{M}_{xy}^{(i)} = m$ for each pair of distinct demes $x^{(i)}$ and $y^{(i)}$ in ancestral population $\mathcal{A}_i$ (Wakeley, 2009). We also assume that all other ancestral populations in the n-taxon tree σ have only one deme (i.e., they are unstructured). At time τ_2, for each $x = 1, 2, \ldots, D$, lineages from deme $x^{(2)}$ in $\mathcal{A}_2$ enter deme $x^{(1)}$ in $\mathcal{A}_1$. At time τ_3, lineages from the λ_{A} subtree enter deme $j^{(2)}$ in $\mathcal{A}_2$ and lineages from λ_{B} enter deme $k^{(2)} \neq j^{(2)}$

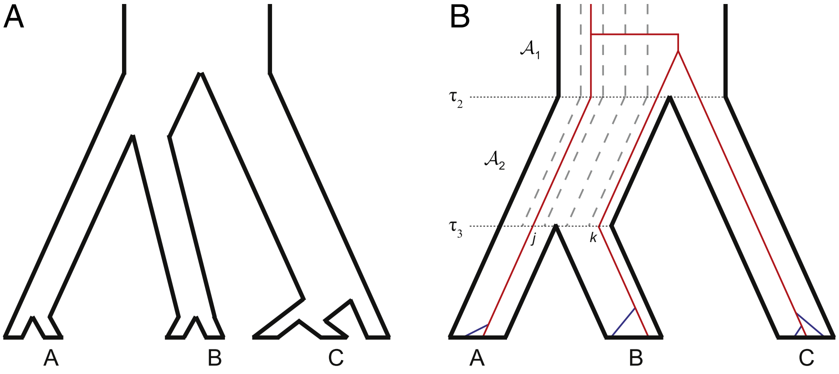


Fig. 2. Example scenario used as a counterexample to prove that consensus methods are misleading. (A) Certain internal branches are made long so that the n-taxon species tree σ resembles a three-taxon species tree. (B) Lineages from species A, B, and C are in red, and lineages from other taxa that have coalesced along the branches leading to species A, B, and C are in blue. The lineage from species A merges into deme $j^{(2)}$ of ancestral population $\mathcal{A}_2$, the lineage from species B merges into deme $k^{(2)} \neq j^{(2)}$ of ancestral population $\mathcal{A}_2$, and the lineage from species C merges into deme $k^{(1)}$ of ancestral population $\mathcal{A}_1$. The migration rates out of demes are small, so that each of the lineages from A, B, and C has a low probability of leaving the deme in which it started. As a consequence, the probability is high that the lineage from B coalesces with the lineage from C before it coalesces with the lineage from A. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in $\mathcal{A}_2$. At time τ_2, lineages from λ_C enter deme $k^{(1)}$, the same deme into which lineages from λ_B, which had entered $k^{(2)}$ in $\mathcal{A}_2$, enter if they have not coalesced or migrated in $\mathcal{A}_2$. We summarize the assumptions in this example scenario as follows.

1. Assumptions about species tree σ
   (a) The species tree σ is fixed and has $n \geq 3$ taxa.
   (b) Certain internal branches on σ are sufficiently long that for a random gene tree $\mathcal{T}$, fixed set of taxa $\mathcal{L}$, where $\mathcal{L}$ is either Γ_A, Γ_B, or Γ_C, and fixed small $\delta > 0$, $\mathbb{P}[\text{top}(\mathcal{T}|\mathcal{L}) = \text{top}(\sigma|\mathcal{L})\ ;\ \mathcal{S}] > 1 - \delta$.
2. Assumptions about the structure of the populations (**D**, **N**, and **M**)
   (a) All populations have one deme except for ancestral populations $\mathcal{A}_1$ and $\mathcal{A}_2$, each of which has a fixed equal number of demes, $D \geq 2$.
   (b) Each deme has a fixed population size of N diploid individuals.
   (c) The population structure model is an island migration model in which the per-generation backward migration rate between each pair of distinct demes within ancestral population $\mathcal{A}_1$ and within ancestral population $\mathcal{A}_2$ is a fixed value m.
3. Assumptions about the species transitions $\boldsymbol{\Psi}$
   (a) At time τ_2, for each $x = 1, 2, \ldots, D$, lineages from deme $x^{(2)}$ in $\mathcal{A}_2$ enter deme $x^{(1)}$ in $\mathcal{A}_1$.
   (b) At time τ_3, lineages from the λ_A subtree enter deme $j^{(2)}$ in $\mathcal{A}_2$, and lineages from the λ_B subtree enter deme $k^{(2)} \neq j^{(2)}$ in $\mathcal{A}_2$.
   (c) At time τ_2, lineages from the λ_C subtree enter deme $k^{(1)}$ in $\mathcal{A}_1$.

In the example scenario of assumptions 1–3, for a specific set of taxa A, B, and C with topological relationship ((AB)C) on σ, we can fix $\tau_2 - \tau_3$, D, and m such that for arbitrarily small $\varepsilon > 0$, the probability that a random gene tree displays the topology ((AB)C) is less than ε. For example, in Fig. 2(B), given fixed $\varepsilon > 0$, fixed $\tau_2 - \tau_3$, and fixed D, for sufficiently small m, with probability greater than $1 - \varepsilon$, the lineage from A and the lineage from B do not migrate, and the lineages from B and C coalesce before either coalesces with the lineage from A. This high probability for coalescence of lineages from B and C causes a large proportion of random gene trees, greater than $1 - \varepsilon$, to display the nonmatching topological relationship ((BC)A).

Define an "event" as either a migration of a lineage from one deme to another deme within an ancestral population or a coalescence of two lineages. Let $p_{\mathcal{S}}(\text{X}, \text{Y})$ be the probability under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$ that a lineage sampled from species X and a lineage sampled from species Y are in the same deme at the speciation time of X and Y. Consider three sampled lineages, one each from species A, B, and C. By construction of the example scenario, $p_{\mathcal{S}}(\text{A}, \text{B}) = 0$ because lineages from A merge into deme $j^{(2)}$ and lineages from B merge into deme $k^{(2)} \neq j^{(2)}$. Within time interval $[\tau_3, \tau_2)$, the time to a migration event in which the lineage from deme $j^{(2)}$ exits the deme is exponentially distributed with rate $(D - 1)m$ per generation and the time to a migration event in which the lineage from deme $k^{(2)}$ exits the deme is also exponential with rate $(D - 1)m$. Therefore, the time to the first migration event (either from deme $j^{(2)}$ or from deme $k^{(2)}$) is exponentially distributed with rate $2(D - 1)m$ per generation (Wakeley, 2009, p. 150, eq. 5.23). Hence, the probability of zero migration events over $[\tau_3, \tau_2)$ – neither for the lineage from A nor for the one from B – is

$$\beta_1 = e^{-2(D-1)m(\tau_2 - \tau_3)}.$$

Treating D and $\tau_2 - \tau_3$ as fixed finite positive values, for sufficiently small migration rate m, β_1 is arbitrarily close to 1. Note, however, that for β_1 to be close to 1, m need not be small—for example, if m is instead a fixed finite positive value and $\tau_2 - \tau_3$ is sufficiently small.

Many possible migration paths exist that can cause a lineage sampled from species B to be located in deme $k^{(1)}$ of population $\mathcal{A}_1$ (the same deme into which lineages from species C merge) at time τ_2. For instance, no migration events might occur or multiple migration events might occur that eventually bring the lineage sampled from species B back into deme $k^{(1)}$ at time τ_2. Because β_1 is the probability of only one among many possible ways for a lineage sampled from species B to be located in deme $k^{(1)}$ at time τ_2, it provides a lower bound for $p_{\mathcal{S}}(\text{B}, \text{C})$. It follows that $\beta_1 < p_{\mathcal{S}}(\text{B}, \text{C}) < 1$, and similarly, a bound can be placed on $p_{\mathcal{S}}(\text{A}, \text{C})$ such that $0 < p_{\mathcal{S}}(\text{A}, \text{C}) < 1 - \beta_1$. Hence, $p_{\mathcal{S}}(\text{A}, \text{C})$ is arbitrarily close to 0 and $p_{\mathcal{S}}(\text{B}, \text{C})$ is arbitrarily close to 1 for sufficiently small m holding $\tau_2 - \tau_3$ fixed, or for sufficiently small $\tau_2 - \tau_3$ holding m fixed. Because the lineages from A and B are in different demes at time τ_3, $p_{\mathcal{S}}(\text{A}, \text{B}) = 0$.

Let $P_{\mathcal{S}}[\mathcal{T}]$ denote the probability that a random gene tree has topology $\mathcal{T}$ under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. If no migration event occurs in the interval $[\tau_3, \tau_2)$, then at time τ_2, the lineage from species A is in deme $j^{(1)}$ and the lineages from species B and C are in deme $k^{(1)} \neq j^{(1)}$. Within the time interval $[\tau_2, \infty)$, the time to the first migration event that causes the lineage from deme $j^{(1)}$ to migrate is exponentially distributed with rate $(D - 1)m$, the time to the first migration event that causes one of the two lineages from deme $k^{(1)}$ to migrate is exponentially distributed with rate $2(D-1)m$, and the time to the event in which the two lineages from deme $k^{(2)}$ coalesce is exponentially distributed with rate $1/(2N)$. Therefore, the time to the first event (migration or coalescence) on the interval $[\tau_2, \infty)$ is exponentially distributed with rate $(D-1)m+2(D-1)m+1/(2N) = 3(D-1)m+1/(2N)$ per generation (Wakeley, 2009, p. 150, eq. 5.23). Hence, the probability that the first event in the interval $[\tau_2, \infty)$ is a coalescence between the lineages from species B and C is $[1/(2N)]/[3(D-1)m+1/(2N)]$. Treating the parameters D and N as fixed finite positive values, the probability that the first event in the interval $[\tau_2, \infty)$ is a coalescence between the lineages from species B and C is arbitrarily close to 1 for sufficiently small m. Multiplying by the probability β_1 of observing zero migration events in the interval $[\tau_3, \tau_2)$, we obtain a lower bound on the probability, β_2, that the first event in the interval $[\tau_3, \infty)$ is a coalescence event between the lineages from species B and C

$$\beta_2 = \frac{1/(2N)}{3(D-1)m+1/(2N)}\beta_1 = \frac{1}{6(D-1)Nm+1}\beta_1. \quad (1)$$

For small enough m, β_2 is arbitrarily close to 1. Note, however, that m need not be too small for a coalescence between lineages from species B and C to be the most probable first event on the interval $[\tau_3, \infty)$. For example, if $\tau_2-\tau_3$ is small, then β_1 is close to 1. Because $P_{\mathcal{S}}[((\text{BC})\text{A})] \geq \beta_2$, for some constant $1/(c+1)$ with $c > 0$, to get $P_{\mathcal{S}}[((\text{BC})\text{A})] > 1/(c+1)$, assuming β_1 is sufficiently close to 1, we would only need $m \approx c/[6(D-1)N]$.

Our example scenario (assumptions 1–3) together with parameter values chosen such that β_2 is arbitrarily close to 1 provides a case in which gene trees have a high probability of containing at least one clade that is not present on species tree σ. A large class of consensus methods then infer species trees with clades not present on σ. This discordance between the inferred species tree topology and σ occurs when σ, $\mathbf{D}$, $\mathbf{N}$, $\mathbf{M}$, and $\boldsymbol{\Psi}$ satisfy assumptions 1–3, when $\tau_2-\tau_3$ and D are fixed, and when m is sufficiently small.

4. Consistency and inconsistency of methods

In this section, under the multispecies coalescent model with ancestral population structure, we investigate the statistical consistency of consensus methods based on seven criteria for inferring species tree topologies. The methods involve using a uniquely favored topology (Democratic Vote), using average coalescence times (STEAC), using average ranks of coalescences (STAR), using uniquely favored rooted triples (R* Consensus and Rooted Triple Consensus), minimizing the number of deep coalescences (MDC), taking the majority rule (Majority-Rule Consensus), and using minimum coalescence times (GLASS/Maximum Tree). We show, through the use of a counterexample, that six of the seven methods are misleading. We also provide a proof that the seventh criterion, based on minimum coalescence times, generates a method that is statistically consistent under our ancestral structure model. The proofs that Democratic Vote is misleading (Proposition 2) and that GLASS/Maximum Tree is consistent (Proposition 11) appear in the main text. The proofs that the other consensus methods are misleading (Propositions 6–8) are similar and are provided in Appendix.

4.1. Uniquely favored topologies

An intuitive approach to inference of species tree topologies is to use Democratic Vote consensus. Democratic Vote estimates a species tree topology using the most frequently occurring gene tree topology, or uniquely favored topology, in a sample of gene trees (Degnan and Rosenberg, 2009). Discordant gene tree topologies that are more probable than the matching topology have been termed "anomalous gene trees" (AGTs), and the space of branch lengths in which AGTs arise has been termed the "anomaly zone" (Degnan and Rosenberg, 2006; Rosenberg, 2013). Owing to AGTs, and because of gene tree discordance more generally, it is difficult for consensus methods to achieve statistical consistency (Degnan et al., 2009). Under the multispecies coalescent model with no ancestral population structure, the space in which Democratic Vote is misleading corresponds exactly to the anomaly zone. A consequence of this correspondence is that Democratic Vote is a statistically consistent estimator of species tree topologies only for three-taxon species trees and for four-taxon species trees with symmetric topologies.

Slatkin and Pollack (2008) showed that for three-taxon species trees, the most likely gene tree topology does not necessarily match the species tree topology under a specific multispecies coalescent model with ancestral population structure. This result implies that in ancestral population structure models, Democratic Vote can be misleading for three-taxon species tree topologies. Our general ancestral structure model, $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$, contains the Slatkin and Pollack (2008) model as a special case. We use the example scenario (assumptions 1–3 in Section 3) as a counterexample to show that Democratic Vote is a misleading estimator for the topology of fixed species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$.

To provide intuition about why Democratic Vote is misleading, assuming that the species tree has topology $((\lambda_A\lambda_B)\lambda_C)$, note that under the example scenario, if we fix $\tau_2-\tau_3$ and D and set the migration rate m sufficiently small, then gene trees generated under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$ display a nonmatching topology with probability arbitrarily close to 1. Because of this large probability, the most frequently occurring gene tree topology—the Democratic Vote topology—is likely to be $((\lambda_B\lambda_C)\lambda_A)$ instead of $((\lambda_A\lambda_B)\lambda_C)$. Thus, Democratic Vote is a misleading estimator for the species tree topology under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$.

Lemma 1. *Consider a species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. Further, consider a random gene tree $\mathcal{T}$. Under assumptions 1–3 in Section 3, for fixed $\delta > 0$,*

$$\begin{aligned}&\mathbb{P}[\text{top}(\mathcal{T}|\Gamma_A) = \text{top}(\sigma|\Gamma_A), \text{top}(\mathcal{T}|\Gamma_B) = \text{top}(\sigma|\Gamma_B),\\ &\quad \text{top}(\mathcal{T}|\Gamma_C) = \text{top}(\sigma|\Gamma_C)\,;\, \mathcal{S}] > 1-\delta.\end{aligned}$$

Proof. Fix a constant $\delta > 0$ and let $\delta' = \delta/3$. Under assumption 1b of the example scenario, given $\delta' > 0$, we can set certain internal branches of σ long enough that for fixed set of taxa $\mathcal{L}$, where $\mathcal{L}$ is either Γ_A, Γ_B, or Γ_C, $\mathbb{P}[\text{top}(\mathcal{T}|\mathcal{L}) = \text{top}(\sigma|\mathcal{L})\,;\, \mathcal{S}] > 1-\delta'$. Using Bonferroni's Inequality (Casella and Berger, 2002, p. 11),

$$\begin{aligned}&\mathbb{P}[\text{top}(\mathcal{T}|\Gamma_A) = \text{top}(\sigma|\Gamma_A), \text{top}(\mathcal{T}|\Gamma_B) = \text{top}(\sigma|\Gamma_B),\\ &\quad \text{top}(\mathcal{T}|\Gamma_C) = \text{top}(\sigma|\Gamma_C)\,;\, \mathcal{S}]\\ &\quad \geq \mathbb{P}[\text{top}(\mathcal{T}|\Gamma_A) = \text{top}(\sigma|\Gamma_A)\,;\, \mathcal{S}]\\ &\qquad + \mathbb{P}[\text{top}(\mathcal{T}|\Gamma_B) = \text{top}(\sigma|\Gamma_B)\,;\, \mathcal{S}]\\ &\qquad + \mathbb{P}[\text{top}(\mathcal{T}|\Gamma_C) = \text{top}(\sigma|\Gamma_C)\,;\, \mathcal{S}] - 2\\ &\quad > (1-\delta') + (1-\delta') + (1-\delta') - 2\\ &\quad = 1-3\delta'\\ &\quad = 1-\delta. \quad \square\end{aligned}$$

Lemma 1 indicates that with a high probability $1-\delta$, for some arbitrarily small $\delta > 0$, a random gene tree will have clades λ_A, λ_B, and λ_C. Consequently, with probability at least $1-\delta$, we will observe gene tree topology $((\lambda_A\lambda_B)\lambda_C)$, $((\lambda_A\lambda_C)\lambda_B)$, or $((\lambda_B\lambda_C)\lambda_A)$. The combined probability of all other topologies is below δ.

Let $\widehat{P}[\mathcal{T}]$ denote the sample proportion of topology $\mathcal{T}$ in a set of L gene trees.

Proposition 2. *Consider a species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. Further, consider a consensus method $\widehat{C}_L$ that estimates* $\text{top}(\sigma)$ *from a set of L gene trees using the most frequently occurring gene tree topology. Then $\widehat{C}_L$ is a misleading estimator of* $\text{top}(\sigma)$.

Proof. We use the example scenario (assumptions 1–3 in Section 3) as a counterexample to show that $\widehat{C}_L$ is misleading. Note that $\text{top}(\sigma) = ((\lambda_A\lambda_B)\lambda_C)$. For $\widehat{C}_L$ to not be misleading, we must have that $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$ as $L \to \infty$. We will instead show that for a different species tree topology $\mathcal{T}^\star = ((\lambda_B\lambda_C)\lambda_A)$, $\widehat{C}_L \xrightarrow{P} \mathcal{T}^\star \neq \text{top}(\sigma)$.

Let $\mathcal{T}$ be a random gene tree. Using Lemma 1, for fixed $\varepsilon > 0$, set migration rate m small enough that

$$\begin{aligned}
P_{\mathcal{S}}[\mathcal{T}^\star] &= P_{\mathcal{S}}[((\lambda_B\lambda_C)\lambda_A)] \\
&> \mathbb{P}[\text{top}(\mathcal{T}|\Gamma_A) = \text{top}(\sigma|\Gamma_A), \text{top}(\mathcal{T}|\Gamma_B) = \text{top}(\sigma|\Gamma_B), \\
&\quad \text{top}(\mathcal{T}|\Gamma_C) = \text{top}(\sigma|\Gamma_C)\,;\, \mathcal{S}]\beta_2 \\
&> (1-\delta)\beta_2 \\
&> (1-\delta)(1-\varepsilon).
\end{aligned}$$

The right-hand side of the last inequality can be made arbitrarily close to 1 for sufficiently small δ and ε. Using the Weak Law of Large Numbers, $\widehat{P}[\mathcal{T}^\star] \xrightarrow{P} P_{\mathcal{S}}[\mathcal{T}^\star]$ as $L \to \infty$. Because $P_{\mathcal{S}}[\mathcal{T}^\star]$ can be made arbitrarily close to 1, $\mathcal{T}^\star$ is the uniquely favored topology and $\mathbb{P}[\widehat{C}_L = \mathcal{T}^\star\,;\, \mathcal{S}] \to 1$ as $L \to \infty$. Therefore, $\widehat{C}_L \xrightarrow{P} \mathcal{T}^\star$, and $\widehat{C}_L$ is a misleading estimator of $\text{top}(\sigma)$. □

4.2. Average distances

Consider a sample of L independent loci. Let d_{XY} be a random variable representing a distance between species X and species Y. We define the estimated mean value of the distance across all L loci,

$$\widehat{d}_{XY} = \frac{1}{L}\sum_{\ell=1}^{L}\widehat{d}_{XY}^{\ell} \tag{2}$$

where $\widehat{d}_{XY}^{\ell}$ is the estimated distance between species X and Y at locus ℓ. If $\mathbb{E}_S[\widehat{d}_{XY}^{\ell}] = \mathbb{E}_S[d_{XY}]$, then by the Weak Law of Large Numbers, the mean distance $\widehat{d}_{XY}$ will be a consistent estimator for the expected distance $\mathbb{E}_{\mathcal{S}}[d_{XY}]$ under model $\mathcal{S} = \mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. We show that in the presence of ancestral structure, methods that employ these consistent mean distances across loci to construct a distance matrix are misleading estimators for the topology of fixed species tree σ with $n \geq 3$ taxa under model $\mathcal{S}$.

To provide intuition about why such a method is misleading, assuming the species tree has topology $((\lambda_A\lambda_B)\lambda_C)$, we can fix $\tau_2 - \tau_3$ and D and set the migration rate m small enough that the lineages from B and C likely coalesce more recently than either coalesces with the lineage from A. A distance measure computed as the mean across loci is likely, with a large number of loci, to have the property that if a pair of lineages is expected to coalesce more recently than another pair, then the pair that is expected to coalesce more recently would have a smaller distance. If the mean sample distances across loci are close enough to the expected distances based on randomly sampled lineages at a random locus, then clustering algorithms applied to the sample distance matrix to reconstruct the species tree topology will construct an incorrect species tree topology rather than the true topology.

Two popular consensus methods that use an average distance across loci are STEAC and STAR (Liu et al., 2009). STEAC employs mean pairwise coalescence times across loci and STAR utilizes average pairwise ranks as distances. We introduce a proposition that applies to both methods as well as others defined by mean distances across loci. For all distance-based methods (including STEAC and STAR), we rely on a result from Jewett and Rosenberg (2012), quoted below as Lemma 4. We first need some definitions.

Definition 3 (*Definition 6.3 of Jewett and Rosenberg, 2012*). Let $b(\sigma)$ denote the length of the shortest branch in a binary species tree σ. Let $\mathbf{d}_\sigma$ be the true matrix of pairwise distances between taxa in tree σ, and let $\widehat{\mathbf{d}}$ be an estimate of $\mathbf{d}_\sigma$. Consider a clustering method $\mathcal{C}$ that takes a distance matrix as input and returns a tree as output. Define the L_∞-norm of matrix $\mathbf{A}$, denoted by $\|\mathbf{A}\|_\infty$, as the magnitude of the largest element in $\mathbf{A}$. The L_∞-radius ℓ_∞ of $\mathcal{C}$ is the supremum over all quantities δ such that, for all species trees σ and all estimates $\widehat{\mathbf{d}}$, $\mathcal{C}$ is guaranteed to return the true topology whenever $\|\widehat{\mathbf{d}} - \mathbf{d}_\sigma\|_\infty < \delta b(\sigma)$.

Lemma 4 (*Proposition 6.4 of Jewett and Rosenberg, 2012*). *Consider a species tree σ, and let $\mathcal{C}$ be a clustering method with nonzero L_∞-radius. Let $\widehat{d}$ be an estimator of the pairwise distance between two species that is consistent as $L \to \infty$. Then the estimator $\widehat{\sigma}$ of the species tree σ produced by applying clustering method $\mathcal{C}$ to the collection $\{\widehat{d}_{XY}\}_{X,Y\in\sigma}$ of distance estimates obtained from $\widehat{d}$ is a consistent estimator for the tree topology as $L \to \infty$.*

Proposition 5. *Consider a species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. Consider a consensus method $\widehat{C}_L$ that estimates* $\text{top}(\sigma)$ *by applying a clustering algorithm satisfying the properties of Lemma 4 to a matrix of statistically consistent pairwise distances across all pairs of taxa, in which the distance $\widehat{d}_{XY}$ between species X and Y is computed as mean distance across L loci as in Eq. (2). Then $\widehat{C}_L$ is a misleading estimator of* $\text{top}(\sigma)$.

4.2.1. Average coalescence times

Consider a sample of L independent loci and let T_{XY}^{ℓ} be a random variable that denotes the coalescence time in the gene tree of locus ℓ for a lineage sampled from species X and a lineage sampled from species Y. Define the average random coalescence time across L loci between one lineage sampled from species X and one lineage sampled from species Y by $\overline{T}_{XY} = (1/L)\sum_{\ell=1}^{L} T_{XY}^{\ell}$. Liu et al. (2009) developed the STEAC consensus method, which utilizes the average coalescence times $\overline{T}_{XY}$, considering each pair of distinct species X and Y, to infer a species tree. The average time $\overline{T}_{XY}$ provides a distance between species X and Y. STEAC creates a distance matrix for all pairs of species and infers a species tree using neighbor-joining. STEAC is a statistically consistent estimator of a species tree topology when gene trees are distributed according to the multispecies coalescent (Liu et al., 2009). This consistency stems from the statistical consistency of neighbor-joining (Atteson, 1999) and the observation that under the multispecies coalescent model, for species X, Y, and Z, if the divergence time of species X and Y, or $d(X, Y)$, is smaller than $d(X, Z)$ and $d(Y, Z)$, then the expected coalescence time is smaller for lineages from X and Y than for lineages from X and Z and for lineages from Y and Z. We show that in the presence of ancestral population structure, STEAC is a misleading estimator for the topology of fixed species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$.

Note that although neighbor-joining is used to construct the STEAC tree from divergence times between pairs of species, a number of other clustering methods can be used to construct species trees from these pairwise times. Lemma 4 states that a species tree estimator that applies a clustering algorithm to the matrix of pairwise divergence times is a statistically consistent estimator of species tree topologies provided that the clustering method has nonzero L_∞-radius, and that the estimator used for divergence times is also a statistically consistent estimator of divergence times. Both neighbor-joining and other common algorithms such as single-linkage clustering, complete-linkage clustering, and UPGMA satisfy Lemma 4 (Jewett and Rosenberg, 2012), and hence Proposition 5 can be applied.

Corollary 6. *Consider a species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. Consider a consensus method $\widehat{C}_L$ that estimates* $\mathrm{top}(\sigma)$ *by applying neighbor-joining to a set of pairwise distances across all pairs of taxa in which the distance between species X and Y is $\widehat{d}_{XY} = 2\overline{T}_{XY}$, twice the average coalescence time over a set of L gene trees for species X and Y. Then $\widehat{C}_L$ is a misleading estimator of* $\mathrm{top}(\sigma)$.

4.2.2. *Average ranks of coalescences*

Coalescence ranks describe the relative order of internal nodes in a rooted tree topology. A ranking is a non-negative real number assignment in which the root is given a positive real value a, and internal nodes are assigned ranks using a function that monotonically decreases with distance along each path from the root to a leaf (Allman et al., 2013; Degnan, 2013). For example, for the topology (((AB)C)D), we could assign a ranking in which the root node has rank 4, the node connecting clade {AB} to C has rank 3, and the node connecting A and B has rank 2. For a symmetric topology ((AB)(CD)), one possible ranking could assign the root node rank 4, and the other two internal nodes rank 3.

In the original algorithm by Liu et al. (2009), the STAR method assumes that the rank of the root node in an n-taxon tree is n. Descending toward the leaves, each internal node is assigned the rank of its immediate ancestor minus 1. Consider a sample of L independent loci, and let R_{XY}^{ℓ} denote the random coalescence rank in the gene tree of locus ℓ for the node that connects a lineage sampled from species X and a lineage sampled from species Y. Denote the random average coalescence rank across L loci between a lineage sampled from species X and a lineage sampled from species Y by $\overline{R}_{XY} = (1/L)\sum_{\ell=1}^{L} R_{XY}^{\ell}$. STAR utilizes the average ranks of coalescences $\overline{R}_{XY}$, for each distinct pair of species X and Y, to infer a species tree. The average rank $\overline{R}_{XY}$ provides a distance between species X and Y. Analogously to the procedure for STEAC, STAR creates a distance matrix for pairs of species and infers a species tree using neighbor-joining. STAR is a consistent estimator of a species tree topology when gene trees are distributed according to the multispecies coalescent (Liu et al., 2009). This consistency stems from the consistency of neighbor-joining and the observation that under the multispecies coalescent, for species X, Y, and Z, if the divergence time of species X and Y is smaller than that for X and Z and for Y and Z, then the expected rank in the gene tree is smaller for the coalescence of lineages from X and Y than for X and Z and for Y and Z. The consistency results still hold with arbitrary rankings in which a non-negative real number is assigned to the root and internal nodes are assigned ranks using a function that monotonically decreases as the number of edges between the node and the root increases (Allman et al., 2013; Degnan, 2013). We show that in the presence of ancestral structure, STAR is a misleading estimator for the topology of fixed species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$.

Corollary 7. *Consider a species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. Consider a consensus method $\widehat{C}_L$ that estimates* $\mathrm{top}(\sigma)$ *by applying neighbor-joining to a set of pairwise distances across all pairs of taxa in which the distance between species X and Y is $\widehat{d}_{XY} = 2\overline{R}_{XY}$, twice the average coalescence rank over a set of L gene trees for species X and Y. Then $\widehat{C}_L$ is a misleading estimator of* $\mathrm{top}(\sigma)$.

4.3. *Majority-rule*

One popular consensus method, Majority-Rule Consensus, constructs a species tree using only clades that appear at frequency greater than some fixed α, $\alpha \in [0.5, 1)$ (Bryant, 2003). The Majority-Rule Consensus tree is either resolved (bifurcating at all nodes), partially unresolved (multifurcating at some nodes), or fully unresolved (multifurcating at the root). For the case of $\alpha = 0.5$, Majority-Rule Consensus has been shown to be a statistically inconsistent, but not misleading, estimator of a species tree topology under the multispecies coalescent (Degnan et al., 2009). We provide a proposition (proven in Appendix) which states that in the presence of ancestral population structure, Majority-Rule Consensus is a misleading estimator for the topology of fixed species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$.

Assuming the species tree has topology $((\lambda_A\lambda_B)\lambda_C)$, by fixing $\tau_2 - \tau_3$ and D and setting the migration rate m small enough that gene trees display topology $((\lambda_B\lambda_C)\lambda_A)$ with probability arbitrarily close to 1, all clades on $((\lambda_B\lambda_C)\lambda_A)$ appear with frequency greater than fixed α. All clades with frequency greater than α appear on the Majority-Rule Consensus tree. Consequently, Majority-Rule Consensus is misleading.

Proposition 8. *Consider a species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. Further, consider a consensus method $\widehat{C}_L$ that estimates* $\mathrm{top}(\sigma)$ *from a set of L gene trees by only using clades present with a frequency greater than fixed α, $\alpha \in [0.5, 1)$. Then $\widehat{C}_L$ is a misleading estimator of* $\mathrm{top}(\sigma)$.

4.4. *Uniquely favored rooted triples*

Define a uniquely favored rooted triple among a set of three taxa X, Y, and Z as the rooted topological relationship among X, Y, and Z with the largest frequency in a sample of rooted gene trees. Because AGTs do not exist for three-taxon species trees under the multispecies coalescent model, consensus methods R* Consensus (Bryant, 2003; Degnan et al., 2009) and Rooted Triple Consensus (Ewing et al., 2008), which infer species trees based on the topologies of uniquely favored rooted triples, have been developed. R* Consensus constructs a species tree from uniquely favored rooted triples through an exact algorithm. Following Degnan et al. (2009), the set $\mathcal{K}$ is a clade in the R* Consensus tree if for each distinct pair of taxa $X', X'' \in \mathcal{K}$ and every taxon $Z \notin \mathcal{K}$, $((X'X'')Z)$ is a uniquely favored rooted triple. The Rooted Triple Consensus tree is constructed by combining the $\binom{n}{3}$ uniquely favored rooted triples using the tree puzzle heuristic (Ewing et al., 2008). Degnan et al. (2009) proved that R* Consensus is a statistically consistent estimator of a species tree topology when gene trees are distributed according to the multispecies coalescent model. We show that in the presence of ancestral population structure, R* Consensus and Rooted Triple Consensus are misleading estimators for the topology of fixed species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. This result follows from the observation that Majority-Rule Consensus is misleading (Proposition 8).

Corollary 9. *Consider a species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. Define the rule $\mathcal{R}$ such that for species X, Y, and Z, X and Y join more recently in the past than do species X and Z and species Y and Z when $\widehat{P}[((XY)Z)] > \widehat{P}[((XZ)Y)]$ and $\widehat{P}[((XY)Z)] > \widehat{P}[((YZ)X)]$. Consider a consensus method $\widehat{C}_L$ that estimates* $\mathrm{top}(\sigma)$ *from a set of L gene trees using uniquely favored rooted triples according to rule $\mathcal{R}$. Then $\widehat{C}_L$ is a misleading estimator of* $\mathrm{top}(\sigma)$.

4.5. Minimizing deep coalescences

Another sensible approach to inferring species trees from gene trees in the presence of incomplete lineage sorting is to minimize the number of deep coalescences (Maddison, 1997). A coalescence event for species X and Y is called "deep" if the event does not occur in the population directly ancestral to the divergence of species X and Y in the species tree. The MDC criterion seeks to find a species tree that minimizes the number of lineages that do not coalesce in the first population in which they have the opportunity to find a common ancestor. Than and Nakhleh (2009) presented an exact method to infer a species tree from gene trees using the MDC criterion. A subsequent study showed that when gene trees are distributed according to the multispecies coalescent model, MDC is a misleading estimator of a species tree topology for four-taxon asymmetric species trees and for species trees with five or more taxa (Than and Rosenberg, 2011). In Appendix, we prove that in the presence of ancestral population structure, MDC is a misleading estimator for the topology of fixed species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$.

Assuming the species tree has topology $((\lambda_A\lambda_B)\lambda_C)$, fix $\tau_2 - \tau_3$ and D and set the migration rate m small enough that gene trees display topology $((\lambda_B\lambda_C)\lambda_A)$ with probability arbitrarily close to 1. The numbers of extra lineages needed to reconcile a gene tree with topology $((\lambda_B\lambda_C)\lambda_A)$ and species trees with topologies $((\lambda_A\lambda_B)\lambda_C)$ and $((\lambda_B\lambda_C)\lambda_A)$ are 1 and 0, respectively. Because the probability of observing a gene tree with topology $((\lambda_B\lambda_C)\lambda_A)$ is high, the species tree with topology $((\lambda_B\lambda_C)\lambda_A)$ minimizes the number of deep coalescences (i.e., the number of extra lineages needed to reconcile the set of gene tree topologies with the species tree topology). Because $((\lambda_B\lambda_C)\lambda_A)$ does not match the species tree topology, MDC is misleading.

Proposition 10. *Consider a species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. Further, consider a consensus method $\widehat{C}_L$ that estimates* top(σ) *from a set of L gene trees by minimizing the number of deep coalescences. Then $\widehat{C}_L$ is a misleading estimator of* top(σ).

4.6. Minimum coalescence times

Consider a sample of L independent loci. Define the minimum coalescence time across L loci between one lineage sampled from species X and one lineage sampled from species Y by $t_{\mathrm{XY}}^{\min} = \min_{\ell=1,\ldots,L} T_{\mathrm{XY}}^{\ell}$. The final method we examine is one that uses $t_{\mathrm{XY}}^{\min}$, considering each distinct pair of species X and Y, to infer a species tree; GLASS (Mossel and Roch, 2010) and Maximum Tree (Liu et al., 2010) are two names for the same method that constructs species trees using these values. The minimum time $t_{\mathrm{XY}}^{\min}$ provides a distance between species X and Y. GLASS/Maximum Tree creates a distance matrix for all pairs of species and infers a species tree using single-linkage clustering. GLASS/Maximum Tree is a statistically consistent estimator of a species tree topology when gene trees are distributed according to the multispecies coalescent (Liu et al., 2010; Mossel and Roch, 2010). In this section, we show that in the presence of ancestral population structure according to our model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$, GLASS/Maximum Tree is a statistically consistent estimator for the topology of fixed species tree σ with $n \geq 3$ taxa.

To provide intuition about why GLASS/Maximum Tree is consistent, note that by assumption, in model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$, demes within an ancestral population communicate through a path of nonzero migration. Because of this communication, as the number of gene trees grows large, for some gene tree a pair of lineages sampled from distinct species likely coalesces in the population directly ancestral to the divergence of those species. Because GLASS/Maximum Tree uses minimum coalescence times to estimate a species tree topology, as the number of gene trees grows large, the single-linkage clustering algorithm applied to these minimum coalescence times yields a tree topology that matches top(σ). Therefore, GLASS/Maximum Tree is a statistically consistent estimator for the species tree topology under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$.

Proposition 11. *Consider a species tree σ with $n \geq 3$ taxa under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. Further, consider a consensus method $\widehat{C}_L$ that estimates* top(σ) *from a set of L gene trees using single-linkage clustering applied to the set of minimum coalescence times $t_{\mathrm{XY}}^{\min}$ for each distinct pair of species X and Y. Then $\widehat{C}_L$ is a statistically consistent estimator of* top(σ).

Proof. This proof is similar to that of Mossel and Roch (2010) for GLASS (see also Jewett and Rosenberg, 2012). Consider L independent loci. For $\widehat{C}_L$ to be consistent, we must have that $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$ as $L \to \infty$. We will show that as the number of loci becomes large, each pair of distinct taxa will, with high probability, have a coalescence event in the population directly ancestral to their split. Further, we will ensure that each of these coalescence events occurs arbitrarily close to the true species split, thereby making each entry in the distance matrix of minimum coalescence times arbitrarily close to the corresponding entry in the distance matrix of divergence times for the true species tree.

Fix arbitrarily small $\varepsilon > 0$, fix the species tree σ, and fix the ancestral population structure model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. Define f_k, $k = 1, 2, \ldots, n-1$, as the probability under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$ that (going back in time) no coalescence occurs between any pair of lineages within ε generations of entering ancestral population A_k. Then over the set of L gene trees, the probability that no coalescence occurs between any pair of lineages within ε generations of entering A_k is $(f_k)^L$. It follows that over the set of L gene trees, the probability that at least one coalescence occurs between each pair of lineages within ε generations of entering A_k is $1 - (f_k)^L$. Therefore, over the set of L gene trees and the set of ancestral populations, the probability that at least one coalescence occurs between each pair of lineages within ε generations of entering each of the $n-1$ ancestral populations is

$$f_{\min} = \prod_{k=1}^{n-1} [1 - (f_k)^L].$$

Because the demes in $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$ communicate, we have $0 < f_k < 1$ for $k = 1, 2, \ldots, n-1$. It follows that $f_{\min} \to 1$ as $L \to \infty$. Consequently, as $L \to \infty$, for each pair of lineages sampled from a pair of species, the minimum coalescence time for those lineages lies within ε generations of the divergence of the two species. Hence, a clustering method satisfying the properties in Lemma 4 that is applied to the set of minimum coalescence times across loci will return the correct tree topology with probability 1 as $L \to \infty$. Single-linkage clustering is such a clustering method (Jewett and Rosenberg, 2012) and hence, applying it to the set of $t_{\mathrm{XY}}^{\min}$ values for all distinct pairs of species X and Y yields top(σ), and $\mathbb{P}[\widehat{C}_L = \text{top}(\sigma)\ ;\ \mathcal{S}] \to 1$ as $L \to \infty$. Therefore, $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$, and $\widehat{C}_L$ is a statistically consistent estimator of top(σ). □

Note that although a method that estimates species tree topologies using minimum coalescence times is statistically consistent, many independent loci might be required for the estimated distance matrix to give rise to an estimate that produces the true species tree. Our proof requires that for each species pair, for at least one locus, a migration must occur immediately ancestral to the species divergence. We can determine the number of loci required for such a migration to be probable.

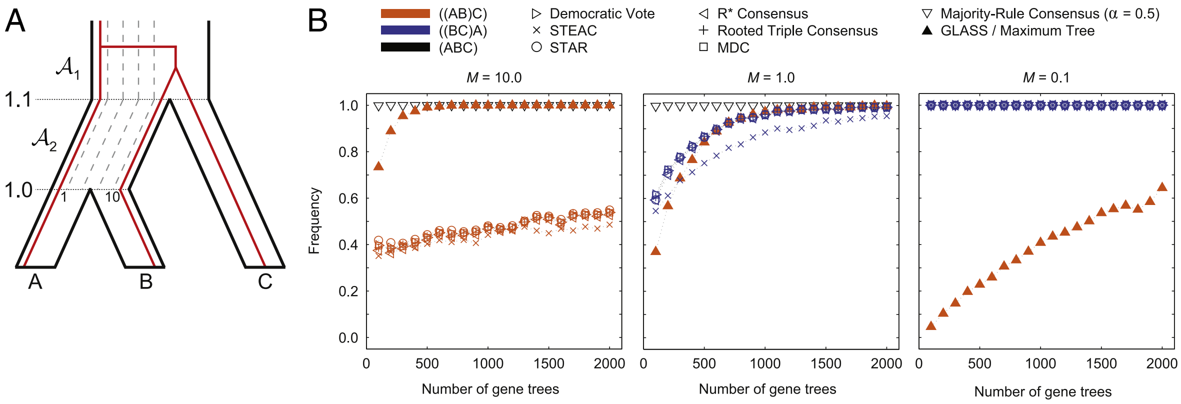


Fig. 3. Simulated probabilities of inferred species trees for the three-taxon model species tree ((AB)C) with ancestral population structure. (A) Species tree with an ancestral population structure model. Time is measured in coalescent units $t/(2N_e)$, where t is time in generations and N_e is a reference diploid effective population size. The population structure model is an island migration model with $D = 10$ demes in each ancestral population. The scaled migration rate between deme x and deme $y \neq x$ is $M = 4N_e m$, which corresponds to $M/4$ individuals per generation in each direction. Species A merges into deme 1 and species B and C each merge into deme 10. (B) Simulation results for scaled migration rates $M = 10.0$, $M = 1.0$, and $M = 0.1$. Each tree topology is represented by a distinct color, and each consensus method is represented by a distinct symbol. For each consensus method, the tree topology traced is the topology for that method that has the highest frequency at 2000 gene trees. The frequency of a topology is calculated as the fraction among 1000 replicate simulations for which that topology was inferred. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

From the example scenario, we know that for a single locus, the probability that no migration occurs within time interval $[\tau_3, \tau_2)$ is $\beta_1 = e^{-2(D-1)m(\tau_2-\tau_3)}$. For a sample of L loci, the probability that no migration occurs within the interval for any of the loci is then $(\beta_1)^L$. Thus, the probability that a migration occurs within the interval for at least one of the loci is $1-(\beta_1)^L$.

To determine the minimum number of loci required for attaining probability at least p that a migration within the interval occurs for at least one sampled locus, we set $1-(\beta_1)^L \geq p$ and solve for L. Taking the logarithm of both sides and solving for the minimum number of loci $L^\star = \min_{L\in\mathbb{Z}^+}\{1-(\beta_1)^L \geq p\}$ yields

$$L^\star = \left\lceil \frac{\log(1-p)}{\log(\beta_1)} \right\rceil = \left\lceil -\frac{\log(1-p)}{2(D-1)m(\tau_2-\tau_3)} \right\rceil.$$

Note that if p is set to a large value or if m or $\tau_2-\tau_3$ is small, then $L^\star$ can be quite large.

5. Simulations

5.1. Comparison of methods on true gene trees

5.1.1. Simulation procedure

To examine the robustness of the eight consensus methods—Democratic Vote, STEAC, STAR, R* Consensus, Rooted Triple Consensus, MDC, Majority-Rule Consensus (with $\alpha = 0.5$), and GLASS/Maximum Tree—to ancestral population structure, we evaluated their performance using simulations. These simulations enabled us to investigate performance on a finite number of loci, rather than in the limiting case. We used the three-taxon species tree $\sigma = ((\text{A}:1.0, \text{B}:1.0):0.1, \text{C}:1.1)$ illustrated in Fig. 3(A). Each of the ancestral populations follows an island migration model with $D = 10$ demes and a scaled unidirectional migration rate between demes $M = 4N_e m$, where N_e is a reference effective number of diploid individuals in a population. Note that because both time and migration rate are scaled by the same effective population size N_e, the specific value of N_e does not matter. Because we are assuming an island migration model, within each ancestral population, for all $i, j \in \{1, 2, \ldots, 10\}$ with $i \neq j$, the migration rate from deme i to deme j is M. Time is measured in coalescent units $t/(2N_e)$, with t measured in generations. Going back in time, the lineage from species A merges into deme 1 in ancestral population $\mathcal{A}_2$, the lineage from species B merges into deme 10 of ancestral population $\mathcal{A}_2$, and the lineage from species C merges into deme 10 of ancestral population $\mathcal{A}_1$. At time τ_2, for each $x = 1, 2, \ldots, 10$, lineages in deme $x^{(2)}$ of $\mathcal{A}_2$ merge into deme $x^{(1)}$ of $\mathcal{A}_1$. This model is precisely the model used for the example scenario in Section 3 with the number of demes set to 10.

Given the species tree model, we used the coalescent simulator MS (Hudson, 2002) to generate gene trees for $L = 100, 200, \ldots, 2000$ independent loci, with each set of L gene trees generated independently of every other set of gene trees. For each consensus method, the L gene trees were then used as input, and a species tree estimate was obtained as output. Each of the consensus methods was applied to the same set of L gene trees. For each L, we repeated this process for 1000 independent replicate simulations of L loci.

5.1.2. Results

Simulation results appear in Fig. 3(B). For scaled migration rate $M = 10.0$, the tree topology with greatest support for each consensus method except for Majority-Rule Consensus is ((AB)C), which matches the species tree. Majority-Rule Consensus instead provides greatest support for the star phylogeny (ABC), reaching frequency 1.0 by 200 gene trees. This result for Majority-Rule applied to three-taxon gene trees is not surprising because Majority-Rule Consensus returns an unresolved three-taxon topology when no input gene tree topology has frequency greater than 0.5. Because the internal branch length is small (0.1 coalescent units) and because the migration rate between demes is large ($M = 10.0$), it is unlikely that any three-taxon gene tree will have frequency greater than 0.5 as the number of input gene trees gets large. The method that performs best is GLASS/Maximum Tree, reaching probability 1 for species tree topology ((AB)C) by 800 gene trees. Although the other six consensus methods provide the strongest support to ((AB)C) at 2000 gene trees, they have low support for ((AB)C), with frequencies of ∼0.55 for STAR, ∼0.54 for Democratic Vote, Rooted Triple Consensus, and MDC, ∼0.53 for R* Consensus, and ∼0.49 for STEAC.

Decreasing the migration rate to $M = 1.0$, we find, as with $M = 10.0$, that GLASS/Maximum Tree has highest support for topology ((AB)C). GLASS/Maximum Tree takes longer than with $M = 10.0$ to converge to the correct topology, reaching frequency 1.0 for ((AB)C) with 1900 instead of 800 genes. As with the $M = 10.0$ case, Majority-Rule Consensus provides greatest support for

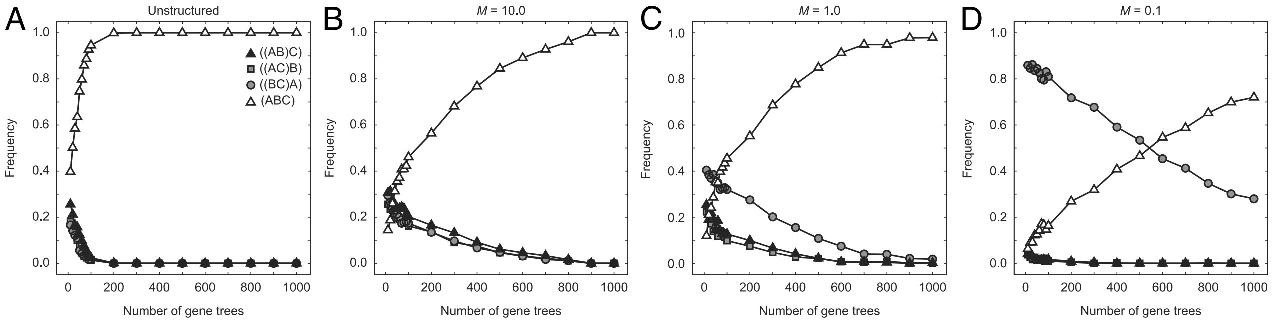


Fig. 4. Inference of species trees using GLASS/Maximum Tree under a Jukes–Cantor substitution model when gene trees are generated under the three-taxon species tree $\sigma = ((\text{A}{:}1.0,\ \text{B}{:}1.0){:}0.1,\ \text{C}{:}1.1)$. The per-site mutation rate θ is 0.01 and time is measured in coalescent units $t/(2N_e)$, where t is time in generations and N_e is a reference diploid effective population size. (A) Simulated probabilities of inferred species trees with no ancestral population structure. (B, C, D) Simulated probabilities of inferred species trees with ancestral population structure. The population structure model is an island migration model with $D = 10$ demes in each ancestral population (as in Fig. 3(A)). The scaled migration rate between deme x and deme $y \neq x$ is $M = 4N_e m = 10.0$ (B), $M = 1.0$ (C), and $M = 0.1$ (D). Species A merges into deme 1 and species B and C each merge into deme 10. Each tree topology is represented by a symbol. The frequency of a topology is calculated as the fraction among 1000 replicate simulations for which that topology was inferred.

(ABC), reaching frequency 1.0 by 200 genes. In contrast to the $M = 10.0$ results, the other six consensus methods no longer have their highest support for the correct topology. Instead, the most favored topology is ((BC)A), reaching frequency ~0.99 at 2000 gene trees for Democratic Vote, STAR, R* Consensus, Rooted Triple Consensus, and MDC and frequency ~0.96 for STEAC. By construction of the simulation, with little enough migration, we expect that each method (except GLASS/Maximum Tree) would infer topology ((BC)A) with highest frequency.

Reducing the migration rate to $M = 0.1$, we find that GLASS/Maximum Tree continues to support the correct species tree topology ((AB)C). Unlike for the two higher migration rates, it does not infer the correct topology with frequency 1.0 by 2000 gene trees, instead obtaining ((AB)C) with frequency ~0.64. However, the frequency of ((AB)C) when inferred by GLASS/Maximum Tree increases as a function of the number of gene trees. Consequently, we expect it would approach 1.0 with enough gene trees, as Proposition 11 predicts. As expected, the other seven consensus methods provide highest support to the topology ((BC)A) with frequency 1.0 for all values of L tested. Majority-Rule gives greatest support for the ((BC)A) topology instead of (ABC) as in the cases for $M = 1.0$ and $M = 10.0$ because when the migration rate is sufficiently small ($M = 0.1$), the probability is far greater than 0.5 that a gene tree displays topology ((BC)A).

5.2. GLASS/Maximum Tree on inferred gene trees with mutation

5.2.1. Simulation procedure

Thus far, our results for GLASS/Maximum Tree have only incorporated genealogical discordance owing to the stochasticity of the coalescent process. However, an additional form of stochasticity that can cause genealogical discordance is mutation. To examine the behavior of GLASS/Maximum Tree when gene trees are estimated instead of known with certainty, we applied GLASS/Maximum Tree to gene trees that were inferred from sequence alignments. We examined the influence of mutation on GLASS/Maximum Tree under two scenarios: with and without ancestral population structure. The species tree in this analysis is identical to that in Section 5.1 and Fig. 3. The only exception is that for the unstructured ancestral population analysis, we let the number of demes in each ancestral population equal 1. To create structured ancestral populations, we use scaled migration rates of $M = 0.1$, 1.0, and 10.0 for the ancestral structure model.

Using MS (Hudson, 2002), we generated gene trees for $L = 10$, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000 independent loci (with each set of L gene trees generated independently of every other set). To convert branch lengths from coalescent units to mutation units (average number of mutations along the branch), we multiplied each length by $\theta/2$, where $\theta = 4N_e\mu = 0.01$ and μ is the mutation rate per site per generation. Each gene tree was input into SEQ-GEN (Rambaut and Grassly, 1997), which generated sequence alignments of length 500 nucleotides under a Jukes–Cantor substitution model. For each alignment, we used PHYLIP (Felsenstein, 1989) to infer rooted gene trees with maximum likelihood assuming the Jukes–Cantor model and a molecular clock. GLASS/Maximum Tree was then applied to the L inferred gene trees. We repeated this process for 1000 independent replicate simulations of L loci.

5.2.2. Results

With no ancestral population structure, as the number of loci increases, GLASS/Maximum Tree is increasingly likely to infer the unresolved phylogeny (ABC) with zero internal branch length (Fig. 4(A)). This increase for (ABC) is caused by maximum likelihood estimating gene trees with branches of length zero due to not having any mutations. Once a single input gene tree has branches of length zero between a pair of species, the GLASS/Maximum Tree estimate must also have branches of length zero between the species pair. As the number of loci increases, the probability increases that the inferred GLASS/Maximum Tree will contain branches of length zero. This increased probability is reflected in the fact that as the number of loci increases, (ABC) increases in frequency and the frequencies of ((AB)C), ((AC)B), and ((BC)A) decrease.

As in the unstructured case, when ancestral populations are structured, the inferred frequency of (ABC) increases with the number of loci (Fig. 4(B, C, D)). However, ancestral structure will generally increase the total branch length of gene trees, decreasing the probability of observing no mutations. This phenomenon decreases the rate at which the frequency of (ABC) increases. As the level of ancestral structure increases (i.e., as migration rate M decreases), the frequency of the ((BC)A) topology becomes higher than that of ((AB)C), for a fixed number of gene trees. This result is likely due to ((AB)C) topologies being converted to (ABC) topologies, because the ((AB)C) topology is expected to have shorter total tree length (and hence fewer mutations) than the ((AC)B) and ((BC)A) topologies.

Finally, because the ancestral structure model in Fig. 4(D) is the same as in Fig. 3 with $M = 0.1$, the probability is small that a gene tree will display topology ((AB)C). This low probability for ((AB)C) is reflected in the small fraction of species trees that have topology ((AB)C) in Fig. 4(D). By incorporating the mutation process in addition to the genealogical process, we find that GLASS/Maximum Tree is increasingly likely to infer an unresolved tree as the number of loci increases.

Table 2
Summary of the behavior of consensus methods.

Criterion	Asymptotic behavior (no structure)	Asymptotic behavior (structure)	Results	Method	Reference
Uniquely favored topologies	Misleading	Misleading	Proposition 2	Democratic vote	Degnan and Rosenberg (2006)
Average coalescence times	Consistent	Misleading	Corollary 6	STEAC	Liu et al. (2009)
Average ranks of coalescences	Consistent	Misleading	Corollary 7	STAR	Liu et al. (2009)
Majority-rule	Inconsistent	Misleading	Proposition 8	Majority-Rule Consensus	Degnan et al. (2009)
Uniquely favored triples	Consistent	Misleading	Corollary 9	R* Consensus	Degnan et al. (2009)
				Rooted Triple Consensus	Ewing et al. (2008)
Minimizing deep coalescences	Misleading	Misleading	Proposition 10	MDC	Than and Nakhleh (2009)
					Than and Rosenberg (2011)
Minimum coalescence times	Consistent	Consistent	Proposition 11	GLASS	Mossel and Roch (2010)
				Maximum Tree	Liu et al. (2010)

An "inconsistent" method does not converge in probability to the correct bifurcating species tree as the number of sampled loci increases; a "misleading" method is not only inconsistent, it converges in probability to an incorrect bifurcating species tree.

6. Discussion

We have described a general ancestral population structure model that extends the basic multispecies coalescent. Using the model, we have proven that many consensus methods for inferring species trees from gene trees that are statistically consistent when ancestral populations are unstructured are no longer consistent when ancestral population structure is introduced (Table 2). The only method that we found to be consistent is GLASS/Maximum Tree, which relies on minimum coalescence times across gene trees between pairs of species. The result, however, does not give a complete perspective on GLASS/Maximum Tree because this method and its extension iGLASS (Jewett and Rosenberg, 2012) have the limitation that if little information exists even in only a single locus in a sample collection of loci, it is possible to obtain an estimated divergence time of 0 between species (Figure 4 and DeGiorgio and Degnan, 2014). Although using the minimum coalescence times between pairs of species is statistically consistent when gene trees are known exactly, the utility of GLASS/Maximum Tree in practice is uncertain.

We note that although our model is more general than the multispecies coalescent, it still provides a simplification of true ancestral population structure. For example, we assume that the migration matrix, the number of demes, and the sizes of the demes are constant along an internal branch of the species tree, and do not account for changes in population size, numbers of demes, and migration rates between demes over time. Even so, this simplified model encodes a counterexample to consistency for many consensus methods.

In addition to methods that are consistent under the standard multispecies coalescent, we considered two methods, Democratic Vote and MDC, that are inconsistent in this basic setting (Degnan and Rosenberg, 2006; Than and Rosenberg, 2011; Rosenberg, 2013). Adding ancestral population structure does not eliminate the inconsistency, and both methods are misleading in the model with ancestral structure. This same reappearance of an inconsistency also applies for Greedy Consensus (Bryant, 2003; Degnan et al., 2009), which refines the tree produced by Majority-Rule Consensus, and which we did not consider in detail. Greedy Consensus is misleading in the standard multispecies coalescent, though Majority-Rule Consensus is only inconsistent and not misleading (Degnan et al., 2009); because we have shown that Majority-Rule Consensus is misleading under ancestral population structure, as a refinement of the Majority-Rule Consensus tree, Greedy Consensus contains all the clades of the Majority-Rule Consensus tree, and therefore continues to be misleading. The relative ease of proving inconsistency in the ancestral structure model for methods found to be inconsistent without ancestral structure illustrates the phenomenon that inconsistency becomes more apparent as new complexities are added to the evolutionary model under investigation.

A number of studies have suggested signatures of ancestral structure in the genomes of a variety of species. For example, White et al. (2009) found significant asymmetry in the distribution of gene trees in a set of three species of mouse, compatible with the type of gene tree incongruence expected from ancestral population structure (Slatkin and Pollack, 2008). Human evolutionary studies have investigated hypotheses regarding ancestral population structure and its effect on gene trees (Innan and Watanabe, 2006; Patterson et al., 2006; Durand et al., 2011; Yamamichi et al., 2012). Yu et al. (2012) examined signatures of ancestral hybridization in yeast by explicitly accounting for discordance that could have arisen from this form of ancestral structure. Continued modeling of the way in which ancestral structure influences the distribution of gene trees will be a useful approach for understanding evolutionary relationships in a variety of taxa.

The observation that many consensus methods are misleading in the presence of ancestral population structure suggests a number of directions for further analysis. First, it will be informative to investigate the effect of ancestral structure on other species tree inference methods that do not follow our modeling framework. Examples of such methods include BEST (Liu and Pearl, 2007; Liu, 2008) and *BEAST (Heled and Drummond, 2010), which perform inference of gene trees and species trees simultaneously rather than applying a consensus algorithm to gene trees; both have been incorporated by Leaché et al. (2014) into species tree inference simulations that involve ancestral structure. It will also be useful to perform additional studies to empirically characterize the properties of ancestral population structure, by taking into account that certain types of discordance, such as asymmetries in the frequencies of gene trees, are signatures of ancestral population subdivision (Slatkin and Pollack, 2008). Results from such studies may be useful in developing consensus methods that are robust to gene tree discordance caused by subdivided ancestral populations.

Acknowledgments

We thank Ethan Jewett, Cuong Than, and two anonymous reviewers for their valuable comments. This work was supported by NIH grant R01 GM117590.

Appendix

In this Appendix, we provide the proofs of results that STEAC, STAR, R* Consensus, Rooted Triple Consensus, Minimize Deep Coalescences, and Majority-Rule Consensus are misleading estimators of a species tree topology under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$. In each case, we use the example scenario described by assumptions

1–3 in Section 3 as a counterexample. For consensus method $\widehat{C}_L$ to not be misleading, we must have that $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$ as $L \to \infty$, where $\text{top}(\sigma) = ((\lambda_A\lambda_B)\lambda_C)$. For all of the methods, the proofs are obtained by considering an alternative species tree topology $\mathcal{T}^\star = ((\lambda_B\lambda_C)\lambda_A)$. We first provide proofs that consensus methods based on mean distances across loci (Proposition 5), such as STEAC (Corollary 6) and STAR (Corollary 7), are misleading. Proofs for Majority-Rule (Proposition 8), R* and Rooted Triple Consensus (Corollary 9), and Minimize Deep Coalescences (Proposition 10) then follow.

Average distances

Under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$, define the expected distance at a random locus for two random lineages, one from species X and the other from species Y, by $\mathbb{E}_{\mathcal{S}}[d_{XY}]$. Let $\mathbb{E}_{\mathcal{S}}[\mathbf{d}]$ be the expected distance matrix under model $\mathcal{S}$, with entries $\mathbb{E}_{\mathcal{S}}[d_{XY}]$. Let d_{XY} be a random variable representing the distance between taxa X and Y, and let $\widehat{\mathbf{d}}$ be the estimated mean distance matrix from a sample of L loci, with entries $\widehat{d}_{XY}$, as in Eq. (2). Recall that $\widehat{d}^{\ell}_{XY}$ is the estimated distance between species X and Y at locus ℓ. Define $\mathbf{d}_{\mathcal{T}}$ as the true distance matrix for topology $\mathcal{T}$, with entry $d_{XY}(\mathcal{T})$ as the distance between species X and Y for topology $\mathcal{T}$.

Proof of Proposition 5. For $\widehat{C}_L$ to not be misleading, we must have that $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$ as $L \to \infty$. Let $\mathcal{T}^\star = ((\lambda_B\lambda_C)\lambda_A)$ be a specific tree topology that differs from the species tree topology $\text{top}(\sigma)$. We will instead show that $\widehat{C}_L \xrightarrow{P} \mathcal{T}^\star \neq \text{top}(\sigma)$ as $L \to \infty$.

For fixed arbitrarily small $\varepsilon > 0$, we must show that $\|\widehat{\mathbf{d}} - \mathbf{d}_{\mathcal{T}^\star}\|_\infty < \varepsilon$ as $L \to \infty$, where $\|\cdot\|_\infty$ is the L_∞-norm. This requirement ensures that a clustering method satisfying the properties in Lemma 4, will, when applied to $\widehat{\mathbf{d}}$, return the incorrect tree topology $\mathcal{T}^\star = ((\lambda_B\lambda_C)\lambda_A)$ with probability 1 as $L \to \infty$.

By the Weak Law of Large Numbers, we have that $\|\widehat{\mathbf{d}} - \mathbf{d}_{\mathcal{T}^\star}\|_\infty \xrightarrow{P} \|\mathbb{E}_{\mathcal{S}}[\mathbf{d}] - \mathbf{d}_{\mathcal{T}^\star}\|_\infty$ as $L \to \infty$, where $\mathbb{E}_{\mathcal{S}}[\mathbf{d}]$ is the expected distance matrix under model $\mathcal{S}$. Define $\mathcal{G}$ to be the set of rooted bifurcating tree topologies on n taxa. It follows that

$$\begin{aligned}
&\|\mathbb{E}_{\mathcal{S}}[\mathbf{d}] - \mathbf{d}_{\mathcal{T}^\star}\|_\infty \\
&\quad= \max_{X,Y} |\mathbb{E}_{\mathcal{S}}[d_{XY}] - d_{XY}(\mathcal{T}^\star)| \\
&\quad= \max_{X,Y} \left| \left(\sum_{\mathcal{T} \in \mathcal{G}} d_{XY}(\mathcal{T}) P_{\mathcal{S}}[\mathcal{T}] \right) - d_{XY}(\mathcal{T}^\star) \right| \\
&\quad= \max_{X,Y} \left| \left(\sum_{\substack{\mathcal{T} \in \mathcal{G} \\ \mathcal{T} \neq \mathcal{T}^\star}} d_{XY}(\mathcal{T}) P_{\mathcal{S}}[\mathcal{T}] \right) + d_{XY}(\mathcal{T}^\star) P_{\mathcal{S}}[\mathcal{T}^\star] - d_{XY}(\mathcal{T}^\star) \right| \\
&\quad= \max_{X,Y} \left| \left(\sum_{\substack{\mathcal{T} \in \mathcal{G} \\ \mathcal{T} \neq \mathcal{T}^\star}} d_{XY}(\mathcal{T}) P_{\mathcal{S}}[\mathcal{T}] \right) - d_{XY}(\mathcal{T}^\star)(1 - P_{\mathcal{S}}[\mathcal{T}^\star]) \right|.
\end{aligned}$$

Define $K = \max_{\mathcal{T}} \{\max_{X,Y} d_{XY}(\mathcal{T})\}$. It follows that $d_{XY}(\mathcal{T}) \leq K$ for any pair of species X and Y, and for any species tree topology $\mathcal{T}$. Fix $\delta > 0$. Then $d_{XY}(\mathcal{T}) < K + \delta$. It follows that

$$\begin{aligned}
&\|\mathbb{E}_{\mathcal{S}}[\mathbf{d}] - \mathbf{d}_{\mathcal{T}^\star}\|_\infty \\
&\quad< \max_{X,Y} \left| \left(\sum_{\substack{\mathcal{T} \in \mathcal{G} \\ \mathcal{T} \neq \mathcal{T}^\star}} (K + \delta) P_{\mathcal{S}}[\mathcal{T}] \right) - d_{XY}(\mathcal{T}^\star)(1 - P_{\mathcal{S}}[\mathcal{T}^\star]) \right| \\
&\quad= \max_{X,Y} |(K + \delta)(1 - P_{\mathcal{S}}[\mathcal{T}^\star]) - d_{XY}(\mathcal{T}^\star)(1 - P_{\mathcal{S}}[\mathcal{T}^\star])| \\
&\quad= \max_{X,Y} |(K + \delta - d_{XY}(\mathcal{T}^\star))(1 - P_{\mathcal{S}}[\mathcal{T}^\star])|.
\end{aligned}$$

By the proof of Proposition 2, we make $P_{\mathcal{S}}[\mathcal{T}^\star]$ arbitrarily close to 1, so that $P_{\mathcal{S}}[\mathcal{T}^\star] > 1 - \varepsilon/(K + \delta)$. Then

$$\begin{aligned}
\|\mathbb{E}_{\mathcal{S}}[\mathbf{d}] - \mathbf{d}_{\mathcal{T}^\star}\|_\infty &< \max_{X,Y} \left| (K + \delta - d_{XY}(\mathcal{T}^\star)) \frac{\epsilon}{K + \delta} \right| \\
&\leq \max_{X,Y} |\epsilon| \\
&= \epsilon.
\end{aligned}$$

Hence, $\|\widehat{\mathbf{d}} - \mathbf{d}_{\mathcal{T}^\star}\|_\infty < \varepsilon$ as $L \to \infty$. By Lemma 4, $\mathbb{P}[\widehat{C}_L = \mathcal{T}^\star\,;\, \mathcal{S}] \to 1$ as $L \to \infty$, and $\mathbb{P}[\widehat{C}_L = \text{top}(\sigma)\,;\, \mathcal{S}] \to 0$. Therefore, $\widehat{C}_L \overset{P}{\not\longrightarrow} \text{top}(\sigma)$, and $\widehat{C}_L$ is a misleading estimator of $\text{top}(\sigma)$. □

Average coalescence times

Under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$, define the expected time to coalescence at a random locus for a random lineage sampled from species X and a random lineage sampled from species Y by $\mathbb{E}_{\mathcal{S}}[T_{XY}]$.

Proof of Corollary 6. For $\widehat{C}_L$ to not be misleading, we must have that $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$ as $L \to \infty$. We will instead show that $\widehat{C}_L \xrightarrow{P} \mathcal{T}^\star \neq \text{top}(\sigma)$ as $L \to \infty$.

By the Weak Law of Large Numbers, $\overline{T}_{XY} \xrightarrow{P} \mathbb{E}_{\mathcal{S}}[T_{XY}]$, and hence the distance $\widehat{d}_{XY} = 2\overline{T}_{XY}$ can be employed by Proposition 5. Thus, by Proposition 5, $\widehat{C}_L$ is a misleading estimator of $\text{top}(\sigma)$. □

Average ranks of coalescences

Let $\mathbb{E}_{\mathcal{S}}[R^{\ell}_{XY}]$ denote the expected rank of a coalescent event for a random lineage from species X and a random lineage from species Y in a gene tree from locus ℓ under $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \boldsymbol{\Psi})$.

Proof of Corollary 7. For $\widehat{C}_L$ to not be misleading, we must have that $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$ as $L \to \infty$. We will instead show that $\widehat{C}_L \xrightarrow{P} \mathcal{T}^\star \neq \text{top}(\sigma)$ as $L \to \infty$.

By the Weak Law of Large Numbers, $\overline{R}_{XY} \xrightarrow{P} \mathbb{E}_{\mathcal{S}}[R_{XY}]$, and hence the distance $\widehat{d}_{XY} = 2\overline{R}_{XY}$ can be employed by Proposition 5. Thus, by Proposition 5, $\widehat{C}_L$ is a misleading estimator of $\text{top}(\sigma)$. □

Majority-rule

Proof of Proposition 8. For $\widehat{C}_L$ to not be misleading, we must have that $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$ as $L \to \infty$. We will instead show that $\widehat{C}_L \xrightarrow{P} \mathcal{T}^\star \neq \text{top}(\sigma)$.

Fix $\alpha \in [0.5, 1)$. Set $\delta > 0$ arbitrarily small and recall from the example scenario that $P_{\mathcal{S}}[\mathcal{T}^\star] = P_{\mathcal{S}}[((\lambda_B\lambda_C)\lambda_A)] > (1 - \delta)\beta_2$, and that when holding $\tau_2 - \tau_3$ and D constant and setting the migration rate m sufficiently small, β_2 is arbitrarily close to 1. For $\widehat{C}_L$ to be misleading, all clades displayed by $\mathcal{T}^\star$ must have frequency greater than α. By the Weak Law of Large Numbers, $\widehat{P}[\mathcal{T}^\star] \xrightarrow{P} P_{\mathcal{S}}[\mathcal{T}^\star]$ as $L \to \infty$. Because $P_{\mathcal{S}}[\mathcal{T}^\star]$ has probability arbitrarily close to 1, all clades displayed by $\text{top}(\mathcal{T}^\star)$ have a frequency greater than α, and $\mathbb{P}[\widehat{C}_L = \mathcal{T}^\star\,;\, \mathcal{S}] \to 1$ as $L \to \infty$. Therefore, $\widehat{C}_L \xrightarrow{P} \mathcal{T}^\star$, and $\widehat{C}_L$ is a misleading estimator of $\text{top}(\sigma)$. □

Uniquely favored rooted triples

Proof of Corollary 9. For $\widehat{C}_L$ to not be misleading, we must have that $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$ as $L \to \infty$. We will instead show that $\widehat{C}_L \xrightarrow{P} \mathcal{T}^\star \neq \text{top}(\sigma)$. From the proof of Proposition 8, all rooted triples displayed by $\mathcal{T}^\star = ((\lambda_B\lambda_C)\lambda_A)$ are uniquely favored. Further, because these uniquely favored triples are derived from a single topology, they are compatible. Because a rooted bifurcating tree topology is defined by its set of rooted triples (Steel, 1992, Proposition 4), $\mathbb{P}[\widehat{C}_L = \mathcal{T}^\star\,;\, \mathcal{S}] \to 1$ as $L \to \infty$. Therefore, $\widehat{C}_L \xrightarrow{P} \mathcal{T}^\star$, and $\widehat{C}_L$ is a misleading estimator of $\text{top}(\sigma)$. □

Minimizing deep coalescences

Consider fixed species tree σ and fixed gene tree topology $\mathcal{T}$ and let $\mathrm{xl}(\mathrm{top}(\sigma), \mathcal{T})$ denote the number of extra lineages contributed by the topology of σ for gene tree topology $\mathcal{T}$. Consider the set of all possible n-taxon rooted bifurcating tree topologies $\mathscr{G}$. The number of extra lineages contributed by the topology of σ is

$$\mathrm{xl}(\mathrm{top}(\sigma)) = \sum_{\mathcal{T} \in \mathscr{G}} \mathrm{xl}(\mathrm{top}(\sigma), \mathcal{T})\widehat{P}[\mathcal{T}]. \tag{3}$$

Proof of Proposition 10. For $\widehat{C}_L$ to not be misleading, we must have that $\widehat{C}_L \xrightarrow{P} \mathrm{top}(\sigma)$ as $L \to \infty$. We will instead show that $\widehat{C}_L \xrightarrow{P} \mathcal{T}^\star \neq \mathrm{top}(\sigma)$. From the example scenario of Section 3, we know that certain branch lengths of the species tree are long enough that for fixed set $X \in \{\Gamma_\mathrm{A}, \Gamma_\mathrm{B}, \Gamma_\mathrm{C}\}$ and fixed arbitrarily small $\delta > 0$, $\mathbb{P}[\mathrm{top}(\mathcal{T}|\mathcal{L}) = \mathrm{top}(\sigma|\mathcal{L})\,;\, \mathscr{S}] > 1 - \delta$. Using Eq. (3), the difference in the numbers of extra lineages contributed by the topologies of σ and $\mathcal{T}^\star$ is

$$\begin{aligned}\Delta_\mathrm{xl}(\mathrm{top}(\sigma), \mathcal{T}^\star) &= \mathrm{xl}(\mathrm{top}(\sigma)) - \mathrm{xl}(\mathcal{T}^\star)\\ &= \sum_{\mathcal{T} \in \mathscr{G}} [\mathrm{xl}(\mathrm{top}(\sigma), \mathcal{T}) - \mathrm{xl}(\mathcal{T}^\star, \mathcal{T})]\widehat{P}[\mathcal{T}].\end{aligned}$$

Note that $\mathrm{xl}(\mathrm{top}(\sigma), \mathcal{T}^\star) = 1$ and $\mathrm{xl}(\mathcal{T}^\star, \mathcal{T}^\star) = 0$. Set the migration rate m small enough that $P_\mathscr{S}[\mathcal{T}^\star] = P_\mathscr{S}[((\lambda_\mathrm{B}\lambda_\mathrm{C})\lambda_\mathrm{A})] > (1 - \delta)\beta_2$, which is arbitrarily close to 1. It follows that, for each $\mathcal{T} \in \mathscr{G} \setminus \{\mathcal{T}^\star\}$, $P_\mathscr{S}[\mathcal{T}]$ is arbitrarily close to 0. For $\widehat{C}_L$ to be misleading, we need $\Delta_\mathrm{xl}(\mathrm{top}(\sigma), \mathcal{T}^\star) > 0$ as $L \to \infty$. For fixed arbitrarily small ε near 0, and by the Weak Law of Large Numbers, Slutsky's Theorem (Serfling, 1980, Theorem 1.5.4), and Corollary B of Serfling (1980), as $L \to \infty$,

$$\begin{aligned}\Delta_\mathrm{xl}(\mathrm{top}(\sigma), \mathcal{T}^\star) &\xrightarrow{P} \sum_{\mathcal{T} \in \mathscr{G}} [\mathrm{xl}(\mathrm{top}(\sigma), \mathcal{T}) - \mathrm{xl}(\mathcal{T}^\star, \mathcal{T})]P_\mathscr{S}[\mathcal{T}]\\ &= [\mathrm{xl}(\mathrm{top}(\sigma), \mathcal{T}^\star) - \mathrm{xl}(\mathcal{T}^\star, \mathcal{T}^\star)]P_\mathscr{S}[\mathcal{T}] + \varepsilon\\ &= P_\mathscr{S}[\mathcal{T}] + \varepsilon\\ &> 0.\end{aligned}$$

Because $\Delta_\mathrm{xl}(\mathrm{top}(\sigma), \mathcal{T}^\star) > 0$ as $L \to \infty$, $\mathbb{P}[\widehat{C}_L = \mathrm{top}(\sigma)\,;\, \mathscr{S}] \to 0$ as $L \to \infty$. Therefore, $\widehat{C}_L \overset{P}{\not\longrightarrow} \mathrm{top}(\sigma)$, and $\widehat{C}_L$ is a misleading estimator of $\mathrm{top}(\sigma)$. □

References

Allman, E.S., Degnan, J.H., Rhodes, J.A., 2013. Species tree inference by the STAR method and its generalizations. J. Comput. Biol. 20, 50–61.

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

Bryant, D., 2003. A classification of consensus methods for phylogenies. In: Janowitz, M., Lapointe, F.J., McMorris, F.R., Mirkin, B., Roberts, F.S. (Eds.), BioConsensus. AMS, Providence, pp. 163–183.

Casella, G., Berger, R.L., 2002. Statistical Inference, second ed. Duxbury, Pacific Grove, CA.

DeGiorgio, M., Degnan, J.H., 2014. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Syst. Biol. 63, 66–82.

Degnan, J.H., 2013. Evaluating variations on the STAR algorithm for relative efficiency and sample size needed to reconstruct species trees. Pac. Symp. Biocomput. 18, 262–272.

Degnan, J.H., DeGiorgio, M., Bryant, D., Rosenberg, N.A., 2009. Properties of consensus methods for estimating species trees from gene trees. Syst. Biol. 58, 35–54.

Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2, e68.

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

Durand, E.Y., Patterson, N., Reich, D., Slatkin, M., 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28, 2239–2252.

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

Felsenstein, J., 1989. PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 5, 164–166.

Garrigan, D., Mobasher, Z., Kingan, S.B., Wilder, J.A., Hammer, M.F., 2005. Deep haplotype divergence and long-range linkage disequilibrium at Xp21.1 provide evidence that humans descend from a structured ancestral population. Genetics 170, 1849–1856.

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

Helmkamp, L.J., Jewett, E.M., Rosenberg, N.A., 2012. Improvements to a class of distance matrix methods for inferring species trees from gene trees. J. Comput. Biol. 19, 632–649.

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

Innan, H., Watanabe, H., 2006. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. Mol. Biol. Evol. 23, 1040–1047.

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

Leaché, A.D., Harris, R.B., Rannala, B., Yang, Z., 2014. The influence of gene flow on species tree estimation: a simulation study. Syst. Biol. 63, 17–30.

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

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

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

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

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

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

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

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, 719–728.

Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S., Reich, D., 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441, 1103–1108.

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238.

Rosenberg, N.A., 2013. Discordance of species trees with their most likely gene trees: a unifying principle. Mol. Biol. Evol. 30, 2709–2713.

Serfling, R.J., 1980. Approximation Theorems of Mathematical Statistics, third ed. Wiley, New York.

Slatkin, M., Pollack, J.L., 2008. Subdivision in an ancestral species creates asymmetry in gene trees. Mol. Biol. Evol. 25, 2241–2246.

Steel, M., 1992. The complexity of reconstructing trees from qualitative characters and subtrees. J. Classification 9, 91–116.

Thalmann, O., Fischer, A., Lankester, F., Pääbo, S., Vigilant, L., 2007. The complex evolutionary history of gorillas: insights from genomic data. Mol. Biol. Evol. 24, 146–158.

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

Than, C.V., Rosenberg, N.A., 2011. Consistency properties of species tree inference by minimizing deep coalescences. J. Comput. Biol. 18, 1–15.

Wakeley, J., 2009. Coalescent Theory: An Introduction. Roberts and Company Publishers, Greenwood Village, Colorado.

White, M.A., Ané, C., Dewey, C.N., Larget, B.R., Payseur, B.A., 2009. Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genet. 5, e1000729.

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

Yamamichi, M., Gojobori, J., Innan, H., 2012. An autosomal analysis gives no genetic evidence for complex speciation of humans and chimpanzees. Mol. Biol. Evol. 29, 145–156.

Yu, Y., Degnan, J.H., Nakhleh, L., 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet. 8, e1002660.