Volume 268
Issue 1
7 January 2011
ISSN 0022-5193
ELSEVIER
Journal of
Theoretical
Biology
50
Years

Journal of Theoretical Biology 289 (2011) 96–106

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

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

Determining species tree topologies from clade probabilities under the coalescent

Elizabeth S. Allman [a], James H. Degnan [b], John A. Rhodes [a,*]

[a] Department of Mathematics and Statistics, University of Alaska Fairbanks, PO Box 756660, Fairbanks, AK 99775, USA
[b] Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

ARTICLE INFO

Article history:
Received 24 February 2011
Received in revised form
8 June 2011
Accepted 2 August 2011
Available online 19 August 2011

Keywords:
Coalescent
Statistical consistency
Identifiability
Consensus
Phylogenetics

ABSTRACT

One approach to estimating a species tree from a collection of gene trees is to first estimate probabilities of clades from the gene trees, and then to construct the species tree from the estimated clade probabilities. While a greedy consensus algorithm, which consecutively accepts the most probable clades compatible with previously accepted clades, can be used for this second stage, this method is known to be statistically inconsistent under the multispecies coalescent model. This raises the question of whether it is theoretically possible to reconstruct the species tree from known probabilities of clades on gene trees.

We investigate clade probabilities arising from the multispecies coalescent model, with an eye toward identifying features of the species tree. Clades on gene trees with probability greater than 1/3 are shown to reflect clades on the species tree, while those with smaller probabilities may not. Linear invariants of clade probabilities are studied both computationally and theoretically, with certain linear invariants giving insight into the clade structure of the species tree. For species trees with generic edge lengths, these invariants can be used to identify the species tree topology. These theoretical results both confirm that clade probabilities contain full information on the species tree topology and suggest future directions of study for developing statistically consistent inference methods from clade frequencies on gene trees.

1. Introduction

A fundamental problem in evolutionary biology is to determine relative relatedness of species, usually by seeking a rooted tree that diagrammatically depicts these relationships. Although phylogenetic methods of inferring relationships between genes sampled from individuals in the different species are now highly developed, such gene trees are not species trees. Even in the absence of errors due to estimating gene trees from DNA sequences, gene tree topologies need not match the underlying species tree. In recent years, various methods have been proposed for inferring species trees from genetic data (Degnan and Rosenberg, 2009; Edwards, 2009; Knowles and Kubatko, 2010). Many of these methods first estimate gene trees, and then resolve the possible conflicts among them to obtain an overall estimate of the species tree.

An important cause of gene tree conflict is the population effect of *incomplete lineage sorting*, in which gene lineages coalesce in ancestral populations earlier than the time these lineages first enter a common ancestral population. The *multispecies coalescent model* (Pamilo and Nei, 1988; Rosenberg, 2002; Rannala and Yang, 2003; Degnan and Salter, 2005; Degnan and Rosenberg, 2009) is commonly used to model this process, producing a distribution of rooted gene trees given a rooted species tree topology and branch lengths (a measure of time and population size on each edge of the species tree). The multispecies coalescent provides a natural framework for incorporating population effects, allowing gene trees to possibly be discordant with the species tree (see Fig. 1), a phenomenon that is very common in multilocus studies (Rokas et al., 2003; Ebersberger et al., 2007; Cranston et al., 2009).

Although the distribution of gene tree topologies from the multispecies coalescent determines the species tree (Allman et al., 2011), estimating this distribution is difficult because there are so many possible topologies: $(2n-3)!!$ when n species are under study. Thus most topologies are unlikely to be observed among a moderate number of gene trees. An alternative is to estimate a smaller set of probabilities which is a function of gene tree probabilities but that still retains enough information to identify the species tree. Other works have considered rooted triples (Degnan et al., 2009; Ewing et al., 2008; Liu et al., 2010) and unrooted gene tree topologies (Allman et al., 2011; Larget et al., 2010). Another possibility, which is

[*] Corresponding author.
E-mail address: j.rhodes@alaska.edu (J.A. Rhodes).

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.08.006

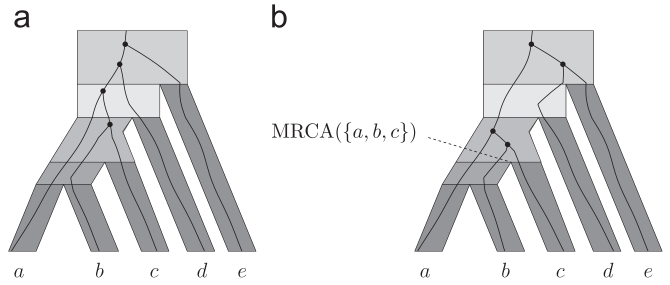

Fig. 1. Gene trees within a species tree. In the multispecies coalescent, gene lineages sampled from species are assumed to coalesce (form nodes in the gene tree) no more recently than their most recent common ancestor (MRCA) in the species tree. Coalescence of lineages in populations more ancient than their MRCA can lead to gene tree topologies that are discordant with the species tree topology. Using upper case letters for gene lineages sampled from their corresponding species, failure of the *A* and *B* lineages to coalesce in their MRCA population makes any of the $\binom{3}{2}$ coalescences between *A*, *B*, and *C* equally likely under the model in the MRCA population of *a*, *b*, and *c*. (a) The gene tree is ((((*B*,*C*),*A*),*D*),*E*). (b) The gene tree is (((*B*,*C*),*A*), (*D*,*E*)).

our focus here, is to use probabilities that a gene tree has a given *clade*, a set of leaves descended from a node of the gene tree that is not ancestral to any other leaves in the gene tree. The probability of a clade under the multispecies coalescent (or any model of gene tree generation) is obtained by simply adding the probabilities of all gene trees that display the given clade (Degnan et al., 2009).

The probability of a clade can be estimated from a collection of gene trees by considering the proportion of gene trees displaying the clade. Since this procedure does not take into account uncertainty in the gene trees, which are themselves estimates from genetic data, a more sophisticated method would quantify the uncertainty in the clades by using posterior probabilities or bootstrap support values for clades obtained from Bayesian or maximum likelihood analyses of the gene trees. The software BUCKy (Ané et al., 2007), for example, takes this approach, using posterior probabilities for clades and additionally incorporating a prior distribution for the amount of gene tree conflict to yield a *concordance factor* for each clade.

One of the most straightforward methods for constructing a species tree from clade probabilities is to use *greedy consensus*, in which the clade with the highest probability (or concordance factor) is accepted, provided it is compatible with previously accepted clades. This process is repeated until a fully resolved tree is formed (Bryant, 2003). This procedure is implemented in BUCKy to construct a *concordance tree*, which is sometimes interpreted as an estimated species tree (Cranston et al., 2009).

To justify a greedy approach, one needs to investigate whether the most probable clades tend also to be clades on the species tree. Indeed, we show in Section 4 that under the multispecies coalescent, any clade with probability greater than 1/3 must be on the species tree, suggesting that the standard majority-rule consensus (which only accepts clades occurring more than 50% of the time) is very conservative in this setting. If the greedy consensus approach is used for clades with probability greater than 1/3 (leaving the tree unresolved with respect to clades with lower probability), then this "not-too-greedy" consensus approach is not misleading, in the sense that it asymptotically cannot return a false species tree clade as the number of loci approaches infinity.

In contrast, previous results have shown that when greedy consensus is applied without restrictions on clade probabilities, the returned tree can be misleading (*i.e.*, for some species trees, as the number of loci increases, the greedy consensus method is increasingly likely to produce a tree that disagrees with the true species tree) for some sets of branch lengths (Degnan et al., 2009). These "too-greedy zones" of edge lengths occur on 4-taxon asymmetric species trees and on any species tree topology with five or more leaves. Thus, caution must be used when probabilities of clades are less than 1/3; it is not obvious how to determine which low-probability clades are on the species tree, even if clade probabilities are known exactly. Other examples show that the most probable k-clade (a clade of $k \geq 2$ elements), is not necessarily a clade on the species tree, even if the species tree is known to have a k-clade.

Undeterred by these negative results, we show in Sections 5 and 6 that under the multispecies coalescent with one lineage sampled per species, the set of clade probabilities does identify the species tree topology for generic branch lengths for any number of species. The proof is based on discovering a linear combination of clade probabilities (a linear invariant) that is equal to zero for any branch lengths on any species tree with a given clade. In theory, if clade probabilities are known, it is therefore possible to identify the species tree by determining all of its clades.

Finally, in Section 6 we extend our results, in part, to cases where the species tree is non-binary and where an arbitrary number of lineages is sampled per species.

Although we frame our questions within the framework of the multispecies coalescent, a careful reading of our arguments reveals that the essential feature of the model that we use is that lineages are *exchangeable*. If two gene lineages are present in the same population at a particular point in time on the species tree, then above that point, the model assumes that both lineages behave the same way. Much of this work, then, should be robust to variations on the coalescent model that preserve exchangeability. Though we do not pursue this here, one could, for instance, consider versions of the multispecies coalescent model in which more than two lineages coalesce simultaneously, as in the Λ-coalescent (Eldon and Wakeley, 2006; Pitman, 1999).

While one might be tempted to use the vanishing of clade invariants for direct inference of clades on a species tree, doing so would require overcoming several obstacles. First, evaluating these invariants on empirical clade probabilities from previously inferred gene trees will rarely yield zero exactly, due to both sampling and gene tree inference errors. Thus it would be necessary to understand the variance of these polynomial values, in order to formulate an appropriate way of determining when values are sufficiently close to zero to indicate a likely clade. Second, the clade invariants we present are not all the constraints on clade probabilities arising from a given species tree. Our clade invariants are all linear equalities, and higher degree equalities can be shown to exist computationally. Moreover, one should expect the existence of non-trivial inequality constraints as well. Ignoring these additional constraints is likely to degrade performance of any such method.

Thus while our linear clade invariants suggest a statistically consistent method of identifying a species tree, how they would perform in practice is unclear. It remains a challenge to incorporate the insight they provide into a practical method that outperforms greedy consensus on most finite data sets. Nonetheless, our results demonstrate that sound statistical inference from clade probabilities is possible.

On a more technical note, there is a key difference in understanding clade probabilities versus many other sets of probabilities related to gene trees or species trees: the failure of marginalization arguments. As this difference plays an important, but unspoken, part throughout this work, we highlight it here.

The problem of establishing identifiability of a species tree from unrooted gene tree probabilities that was taken up previously (Allman et al., 2011) is superficially similar to the clade problem of this paper. Both unrooted gene tree probabilities and clade probabilities can be obtained by summing probabilities of appropriate rooted gene trees. The sum is either over all rooted gene trees with the same unrooted topology or over all rooted gene trees that have the clade in question.

Note also that the probability of a gene tree on a subset of the taxa can be obtained by summing probabilities of gene trees on the full set that display the given gene tree when restricted to the subset. As an example, if there are four lineages, A, B, C, and D, the probability that a gene tree restricted to lineages A, C, and D has the topology $((C,D),A)$ can be obtained by marginalizing over B:

$$\begin{aligned}\mathbb{P}_{\sigma_R}[((C,D),A)] &= \mathbb{P}_{\sigma}[(((C,D),A),B)]+\mathbb{P}_{\sigma}[(((C,D),B),A)]\\ &\quad+\mathbb{P}_{\sigma}[(((B,C),D,),A)]+\mathbb{P}_{\sigma}[(((B,D),C),A)]\\ &\quad+\mathbb{P}_{\sigma}[((A,B),(C,D))], \end{aligned}\tag{1}$$

where σ is the species tree on all four taxa, and σ_R is a reduced species tree obtained by removing the species with lineage B. We can therefore think of the probability of the gene tree on the reduced set of taxa as a linear combination of the gene tree probabilities on the full set of taxa (where the coefficients of the linear combination are either 0 or 1). Moreover this marginalization formula is independent of the topology of the 4-taxon species tree.

Such marginalization of a gene tree distribution to fewer taxa is possible for either rooted or unrooted gene trees. Consequently, for most arguments in Allman et al. (2011) it was sufficient to focus on small trees, with at most five taxa. Indeed, similar marginalization arguments are standard throughout phylogenetic theory.

Unfortunately, a marginalization approach fails for studying clades when the species tree is unknown. Given clade probabilities arising from an n-taxon species tree σ under the multispecies coalescent, one would like to be able to determine clade probabilities arising from an induced k-taxon tree displayed on σ. However, probabilities of clades on the k-taxon induced tree cannot be obtained from a linear combination of the clade probabilities associated with the n-taxon species tree without knowledge of the species tree. That is, for clades there is no linear formula analogous to (1) which is independent of the species tree. We demonstrate this formally in the case where $k=3$ and $n=4$ in Appendix A.

This inability to marginalize clade probabilities without knowing the species tree topology motivated looking for an invariant that would hold for clades on trees of any size. Although only linear invariants are needed in the proof of identifiability, the invariants constructed for k-clades involve a linear combination of 2^{k-1} clade probabilities. These rather elaborate invariants and the inability to marginalize clade probabilities to smaller trees lead to a different flavor for the proof of species tree identifiability from clade probabilities.

2. Definitions

Let $\mathcal{X}$ be a finite set, whose elements we refer to as *taxa*. A *species tree on* $\mathcal{X}$ means a pair $\sigma=(\psi,\lambda)$, where ψ is a rooted, topological tree whose leaves are bijectively labelled by elements of $\mathcal{X}$, and $\lambda=(\lambda_1,\dots,\lambda_k)$ is a collection of lengths for the internal branches of ψ. We refer to ψ as a *species tree topology*, and always assume all internal nodes of ψ except the root have degree at least 3. If all internal nodes except the root have degree 3 and the root has degree 2, we say that ψ and σ are *binary*.

We use a modified Newick notation for species trees, as in Allman et al. (2011), in which we do not specify the lengths of pendant edges, since only the lengths of internal edges affect probabilities of gene tree topologies under the multispecies coalescent. For example, we write $((a,b):t,c)$ for a 3-taxon species tree with one internal edge with length t, measured in coalescent units. If there is a constant effective population size, N, over an edge of the species tree, then a length of t indicates that the edge represents Nt generations (Degnan and Rosenberg, 2009). For varying effective population size, a non-linear scaling is needed to relate coalescent units to generations. Species trees are thus not assumed to be ultrametric in coalescent units.

In discussing trees, we find it convenient in various settings to use either spatial or temporal terminology. For instance, if (v, w) is a directed edge in ψ pointing away from the root, then we may say that v is *above*, or an *ancestor*, of w and that w is *below*, or a *descendant*, of v. Natural extensions of these terms should be clear from context.

We denote taxa in $\mathcal{X}$ by lower case letters such as $a,b,c,\dots$. To distinguish between taxa and sampled genes from those taxa, we use the corresponding upper case letters $A,B,C,\dots$ to denote the genes, with the set $\mathcal{X}_g$ denoting the full set of genes, one for each taxon. Similarly, a subset of taxa $\mathcal{C}\subseteq\mathcal{X}$ has a corresponding subset of genes $\mathcal{C}_g\subseteq\mathcal{X}_g$. A sampled gene tree from the multispecies coalescent model on σ will thus have leaves labelled by $\mathcal{X}_g$, and in general may have any topology, regardless of the species tree topology ψ. More specifically, by a *gene tree T* we mean a binary, rooted topological tree with leaves bijectively labelled by $\mathcal{X}_g$. We emphasize that for this article gene trees are topological only, with no edge lengths specified. We require that gene trees be binary, since under the multispecies coalescent only binary gene trees have positive probability.

Definition. If ψ is a species tree topology on $\mathcal{X}$, and $\mathcal{A}\subseteq\mathcal{X}$, then the *most recent common ancestor of* $\mathcal{A}$, MRCA($\mathcal{A}$), is the node of ψ that is ancestral to all elements of $\mathcal{A}$ and which is a descendant of any other node ancestral to all elements of $\mathcal{A}$.

Definition. Let $\mathrm{desc}_{\mathcal{X}}(v)\subseteq\mathcal{X}$ denote the elements of $\mathcal{X}$ descended from a node v of a species tree topology ψ on $\mathcal{X}$, so that if $\mathcal{A}\subseteq\mathcal{X}$, then $\mathcal{A}\subseteq\mathrm{desc}_{\mathcal{X}}(\mathrm{MRCA}(\mathcal{A}))\subseteq\mathcal{X}$. A *clade* $\mathcal{C}$ on ψ is a subset of $\mathcal{X}$ such that $\mathcal{C}=\mathrm{desc}_{\mathcal{X}}(\mathrm{MRCA}(\mathcal{C}))$.

The notions of MRCA and clade extend to gene trees in an obvious way, replacing $\mathcal{X}$, ψ, $\mathcal{C}$, with $\mathcal{X}_g$, T, $\mathcal{C}_g$ in the definitions.

Definition. For a gene tree T, the set of all clades on T is denoted $\mathcal{H}(T)$. Similarly, for a species tree $\sigma=(\psi,\lambda)$ the set of clades on ψ is denoted $\mathcal{H}(\sigma)=\mathcal{H}(\psi)$.

In discussing the relationships between a subset $\mathcal{Y}$ of the taxa $\mathcal{X}$ on a tree ψ, we use the terminology of a *displayed tree*: a tree obtained from the full tree by first passing to the rooted subtree spanned by $\mathcal{Y}$, and then suppressing any non-root nodes of degree 2 (Semple and Steel, 2003). As an example, the species tree in Fig. 1 displays $((b,d),e)$. Such 3-taxon trees displayed within a larger tree are also called *rooted triples*. The notion of displayed trees can be applied in the context of either species trees (with or without branch lengths) or gene trees.

A detailed presentation of the multispecies coalescent model has been given previously (Allman et al., 2011), so we omit repeating that here. Because we focus in this paper on the probabilities of observing gene trees or clades on gene trees under that model, we fix the following notation.

Definition. Under the multispecies coalescent model on a fixed species tree σ on taxa $\mathcal{X}$, the probabilities of a gene tree T, and a clade $\mathcal{C}_g$ on gene trees are denoted $\mathbb{P}_{\sigma}(T)$ and $\mathbb{P}_{\sigma}(\mathcal{C}_g)$, respectively.

If more than one lineage is sampled per species, a generalization of our results on species tree identifiability still holds. For this extension, we require the following definitions. (See Fig. 2 for an example.)

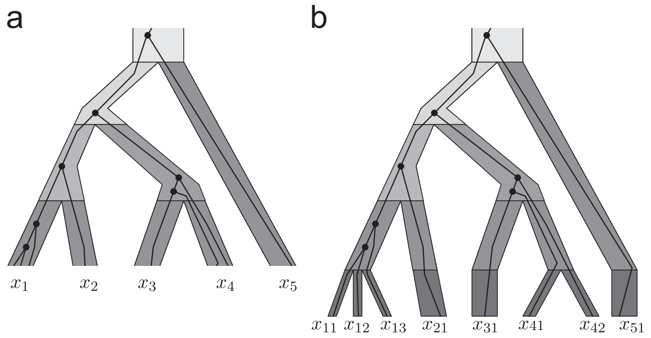

Fig. 2. (a) A gene tree with multiple lineages sampled from several species within a species tree. The taxa are $\mathcal{X}=\{x_1,x_2,x_3,x_4,x_5\}$ with $\delta=(3,1,1,2,1)$ lineages sampled from them. (b) The extended version of the species tree with taxa $\mathcal{X}^*=\{x_{11},x_{12},\ldots,x_{51}\}$ and one lineage sampled per taxon. Under the multispecies coalescent, the probability of any clade $\mathcal{A}_g\subset\mathcal{X}_g^*$ is the same for both the species trees in (a) and in (b).

Definition. Let $\mathcal{X}=\{x_1,\ldots,x_n\}$ be a taxon set, with $|\mathcal{X}|=n$. Let $\delta=(\delta_1,\ldots,\delta_n)$ be the number of individuals sampled from species x_i, $i=1,\ldots,n$. With x_{ij}, $1\le j\le\delta_i$, denoting the individuals in taxon x_i, $\mathcal{X}^*=\{x_{ij}\}$ is the set of all sampled individuals, so $|\mathcal{X}^*|=\sum_{i=1}^n\delta_i$.

An *extended species tree* $\sigma^*=(\psi^*,\lambda,\delta)$ on $\mathcal{X}$ is a species tree (ψ^*,λ) on $\mathcal{X}^*$ such that for each $1\le i\le n$ all the leaves x_{ij}, $1\le j\le\delta_i$ have a common parent in ψ^*.

The *pruned species tree topology* ψ on $\mathcal{X}$ is obtained from ψ^* by labelling the parent of the x_{ij} by x_i for each i with $\delta_i>1$, and then excising the leaves x_{ij} and the pendant edges on which they lie.

Note that while an extended species tree gives rise to a species tree by the pruning process, in an extended species tree a branch length is assigned to those edges which become pendant in the species tree whenever there are two or more sampled individuals in the taxon. Since our notion of a species tree in this paper does not have pendant edge lengths, an extended species tree thus carries more edge length information than the associated species tree.

Gene trees arising from the coalescent model on an extended species tree have leaves labelled $\mathcal{X}_g=\{X_{11},\ldots,X_{1\delta_1},\ldots,X_{n1},\ldots,X_{n\delta_n}\}$ and are, with probability 1, binary. One readily checks that the probability of such a gene tree under the multispecies coalescent on the extended species tree is exactly the same as the probability of the gene tree under a multiple individual sampling scheme on the species tree (with some pendant edge lengths) obtained by pruning. Indeed, this is why we have introduced such trees. We will use them to easily extend results where one individual is sampled per species to the multiple sampling situation, in Proposition 13 and Corollary 14.

Finally, note that by construction, for each $i=1,\ldots,n$, the set $\mathcal{A}_i=\{x_{i1},\ldots,x_{i\delta_i}\}$ is a clade on the extended species tree. But of course a set $(\mathcal{A}_i)_g=\{X_{i1},\ldots,X_{i\delta_i}\}$ need not be a clade on any given gene tree.

3. Arbitrary gene tree distributions

Though the remainder of this paper is concerned only with the gene tree distribution arising from the multispecies coalescent model, in this section we investigate clade probabilities for arbitrary binary gene tree distributions. The main observation is that without special assumptions on the gene tree distribution, the clade probabilities do not contain enough information to recover the gene tree distribution.

Note that every gene tree must have as clades all singleton sets of gene labels, as well as the full set $\mathcal{X}_g$. We refer to these as *trivial clades*. Any other nonempty subset $\mathcal{C}_g\subset\mathcal{X}_g$ is a clade on some gene trees, but not others.

For an arbitrary distribution of gene trees on a taxon set $\mathcal{X}$, let $\mathbb{P}(T)$ denote the probability of gene tree T. Then for each subset $\mathcal{C}_g\subseteq\mathcal{X}_g$, the probability that $\mathcal{C}_g$ is a clade on a gene tree is

$$\mathbb{P}(\mathcal{C}_g)=\sum_T\mathbb{P}(\mathcal{C}_g|T)\mathbb{P}(T)=\sum_T I(\mathcal{C}_g\in\mathcal{H}(T))\mathbb{P}(T),$$

where I is the indicator function with values of 1 or 0. Note that the probability of any trivial clade is therefore 1.

We emphasize that the clade probabilities for an n-taxon species tree σ do *not* form a probability distribution. The presence of different clades may not be mutually exclusive events (for instance, if $\mathcal{C}_g\subset\mathcal{C}_g'$), and their probabilities do not sum to 1.

Proposition 1. *If $|\mathcal{X}|=n$, then for any distribution of binary gene trees the sum of the probabilities of all non-trivial clades is $n-2$.*

Proof. Denoting n-taxon gene trees by T,

$$\sum_{\substack{\mathcal{C}_g\subset\mathcal{X}_g\\ \text{non-trivial}}}\mathbb{P}(\mathcal{C}_g)=\sum_{\substack{\mathcal{C}_g\subset\mathcal{X}_g\\ \text{non-trivial}}}\sum_T I(\mathcal{C}_g\in\mathcal{H}(T))\mathbb{P}(T)=\sum_T\sum_{\substack{\mathcal{C}_g\subset\mathcal{X}_g\\ \text{non-trivial}}}I(\mathcal{C}_g\in\mathcal{H}(T))\mathbb{P}(T).$$

But since each binary gene tree has $n-2$ non-trivial clades, this shows that

$$\sum_{\substack{\mathcal{C}_g\subset\mathcal{X}_g\\ \text{non-trivial}}}\mathbb{P}(\mathcal{C}_g)=\sum_T(n-2)\mathbb{P}(T)=n-2.\qquad\square\qquad(2)$$

Theorem 2. *For an arbitrary distribution of binary gene trees on a taxon set $\mathcal{X}$ with $|\mathcal{X}|\ge4$, the gene tree probabilities $\mathbb{P}(T)$ cannot be identified from the clade probabilities $\mathbb{P}(\mathcal{C}_g)$.*

Proof. The set $\mathcal{X}_g$ has 2^n-n-2 subsets $\mathcal{C}_g$ with $2\le|\mathcal{C}_g|\le n-1$. Using Proposition 1, the clade probabilities can thus be specified by a point in a (2^n-n-3)-dimensional vector space. However, there are $(2n-3)!!=1\cdot3\cdots(2n-3)$ binary gene trees on $\mathcal{X}_g$, so a gene tree distribution is specified by a point in a $((2n-3)!!-1)$-dimensional vector space. But since

$$(2n-3)!!-1>2^n-n-3$$

when $n\ge4$, and the map from gene tree probabilities to clade probabilities is linear, the map is not invertible at any point. □

We note that for an arbitrary distribution on multifurcating gene tree topologies, the trivial invariant in Eq. (2) need not hold. However, the argument establishing Theorem 2 can be modified to apply to such distributions, since the number of multifurcating trees is greater than the number of binary ones.

4. Highly probable gene tree clades are species tree clades

For the remainder of the paper, we assume that both gene tree probabilities $\mathbb{P}_\sigma(T)$ and clade probabilities $\mathbb{P}_\sigma(\mathcal{C}_g)$ arise from the multispecies coalescent on a species tree $\sigma=(\psi,\lambda)$.

Theorem 3. *Let $\sigma=(\psi,\lambda)$ be a binary species tree on $\mathcal{X}$, with edge lengths $\lambda_i>\varepsilon\ge0$. Under the multispecies coalescent model, suppose $\mathcal{C}_g\subset\mathcal{X}_g$ has clade probability $\mathbb{P}_\sigma(\mathcal{C}_g)\ge(1/3)\exp(-\varepsilon)$. Then $\mathcal{C}$ is a clade on σ; that is, $\mathcal{C}\in\mathcal{H}(\sigma)$.*

Furthermore, if $(1/3)\exp(-\varepsilon)$ is replaced with any smaller number, this statement is no longer true for all such choices of species trees and non-trivial clades: For any $k<(1/3)\exp(-\varepsilon)$, there exists a species tree σ on $\mathcal{X}$ and a taxon set $\mathcal{C}\subset\mathcal{X}$ with $1<|\mathcal{C}|<|\mathcal{X}|$ such that $\mathcal{C}$ is not a clade on σ, yet $\mathbb{P}_\sigma(\mathcal{C}_g)\ge k$.

Proof. If $\mathcal{C}$ is a trivial clade, there is nothing to show, so we may assume $1<|\mathcal{C}_g|<|\mathcal{X}|$. We prove the contrapositive: if $\mathcal{C}$ is not a clade on ψ, then $\mathbb{P}_\sigma(\mathcal{C}_g)<(1/3)\exp(-\varepsilon)$.

Suppose $\mathcal{C}$ is not a clade on the species tree, so there exist $a,b \in \mathcal{C}$ and $c \in \mathcal{X}\backslash\mathcal{C}$ such that ψ does not display the rooted triple $((a,b),c)$. Thus, the rooted triple probability satisfies $\mathbb{P}_\sigma(((A,B),C)) < (1/3)\exp(-\varepsilon)$ (Nei, 1987). But then

$$(1/3)\exp(-\varepsilon) > \mathbb{P}_\sigma(((A,B),C)) \geq \mathbb{P}_\sigma(\mathcal{C}_g),$$

since $((A,B),C)$ is displayed on every gene tree on which $\mathcal{C}_g$ is a clade.

To establish the last claim of the theorem, we construct an example. For any set $\mathcal{C}$ with $1 < |\mathcal{C}| < |\mathcal{X}|$, pick some $a \in \mathcal{C}$, and some $c \in \mathcal{X}\backslash\mathcal{C}$. Let $\mathcal{C}' = \mathcal{C}/\{a\}$. Consider a binary species tree σ which has a subtree of the form $((a,c):\delta,T_{\mathcal{C}'}:\gamma)$, where $T'_{\mathcal{C}}$ is any rooted tree on $\mathcal{C}'$. Note then that $\mathcal{C}$ is not a clade on σ.

By taking γ to be large, the probability that the lineages from $\mathcal{C}'$ coalesce below MRCA$(\{a,c\} \cup \mathcal{C}')$ can be made as close to 1 as desired. Because the probability that lineages A and C fail to coalesce within time δ is $\exp(-\delta)$, by also choosing $\delta \approx \varepsilon$ the probability that three lineages (one for A, one for C, and one for $\mathcal{C}'_g$) enter the ancestral population above this MRCA can be made as close to $\exp(-\varepsilon)$ as we wish. Thus the probability that $\mathcal{C}_g$ will be a clade on a gene tree can be made as close to $(1/3)\exp(-\varepsilon)$ as we wish. □

Setting $\varepsilon = 0$ yields Corollary 4.

Corollary 4. *Let σ be a binary species tree on taxa $\mathcal{X}$, with positive edge lengths. Under the multispecies coalescent model, suppose $\mathcal{C} \subset \mathcal{X}$ is such that $\mathbb{P}_\sigma(\mathcal{C}_g) \geq 1/3$. Then $\mathcal{C}$ is a clade on σ.*

Furthermore, this statement is no longer true for $1 < |\mathcal{C}| < |\mathcal{X}|$ if 1/3 is replaced with any smaller number.

If the species tree is not binary, a slightly weaker result holds, requiring a strict lower bound on the clade probability.

Theorem 5. *Suppose the species tree σ is not necessarily binary, and $\mathcal{C} \subset \mathcal{X}$ is such that $\mathbb{P}_\sigma(\mathcal{C}_g) > 1/3$. Then $\mathcal{C}$ is a clade on σ.*

Furthermore, this statement is no longer true for $1 < |\mathcal{C}| < |\mathcal{X}|$ if 1/3 is replaced with any smaller number.

Proof. To show $\mathcal{C}$ is a clade, we suppose $c \in \mathcal{X}\backslash\mathcal{C}$ and demonstrate that $c \notin \mathrm{desc}_{\mathcal{X}}(\mathrm{MRCA}(\mathcal{C}))$. Choose $a,b \in \mathcal{C}$ such that $\mathrm{MRCA}(\mathcal{C}) = \mathrm{MRCA}(\{a,b\})$. Note that $\mathbb{P}_\sigma(((A,B),C)) \geq \mathbb{P}_\sigma(\mathcal{C}_g)$ since any gene tree displaying the clade $\mathcal{C}_g$ must display the rooted triple $((A,B),C)$. This implies $\mathbb{P}_\sigma(((A,B),C)) > 1/3$ and thus that the rooted triple $((a,b),c)$ is displayed on σ. Thus $c \notin \mathrm{desc}_{\mathcal{X}}(\mathrm{MRCA}(\mathcal{C}))$.

That 1/3 cannot be replaced with a smaller number is a consequence of Corollary 4. □

5. Clade invariants

A *clade invariant* for a species tree topology is a polynomial in the probabilities of clades on gene trees that vanishes for all edge length assignments to the species tree. More completely, a clade invariant associated to an n-taxon species tree topology ψ is a multivariate polynomial in 2^n-n-2 indeterminates (one for every non-trivial clade) which evaluates to zero at any vector of clade probabilities $\mathbb{P}_\sigma(\mathcal{C}_g)$ arising from $\sigma = (\psi,\lambda)$, regardless of the values of λ.

Proposition 1 gives an example of a clade invariant for binary gene trees that, in addition, is independent of all features of ψ except the number of taxa:

$$\sum_{\substack{\mathcal{C}_g \subset \mathcal{X}_g \\ \text{non-trivial}}} \mathbb{P}_\sigma(\mathcal{C}_g) - (n-2) = 0.$$

We call this the *trivial invariant*, and emphasize that it is satisfied by clade probabilities from any species tree on $\mathcal{X}$.

Clade invariants can be computed for small trees using computational algebra software, such as Singular (Greuel et al., 2009). For each edge length λ_i, one sets $\Lambda_i = \exp(-\lambda_i)$, and then expresses the clade probabilities as multivariate polynomials in the Λ_i. Gröbner basis methods for variable elimination then allow one to determine generators of the polynomial ideal of all clade invariants. Such computations were useful in formulating the general construction of certain linear invariants given below. The existence of these clade invariants forms the basis for our proof of species tree topology identifiability in Section 6.

Theorem 6. *Let $\mathcal{A} \subsetneq \mathcal{X}$ be a subset of taxa with at least two elements, and $\mathcal{C} \subseteq \mathcal{X}\backslash\mathcal{A}$ a non-empty set of taxa not in $\mathcal{A}$. For distinct $a,b \in \mathcal{A}$, let $\mathcal{A}' = \mathcal{A}\backslash\{a,b\}$. Then if $\mathcal{A}$ is a clade on σ,*

$$\left(\sum_{\mathcal{S} \subseteq \mathcal{A}'} \mathbb{P}_\sigma(\mathcal{S}_g \cup \{A\} \cup \mathcal{C}_g)\right) - \left(\sum_{\mathcal{S} \subseteq \mathcal{A}'} \mathbb{P}_\sigma(\mathcal{S}_g \cup \{B\} \cup \mathcal{C}_g)\right) = 0. \tag{3}$$

We note that this theorem applies to any species tree, including non-binary ones. Moreover, since a non-binary species tree σ can be thought of as any of its binary resolutions with length 0 assigned to any introduced edges, the clade probabilities arising from such a σ will satisfy the polynomials of the theorem for every binary resolution. Thus in the statement of the theorem the phrase 'if $\mathcal{A}$ is a clade on σ' can be replaced with 'if $\mathcal{A}$ is a clade on a binary resolution of σ.'

For the proof, it is useful to have the notion of compatible clades:

Definition. Two clades, $\mathcal{A}_g$ and $\mathcal{B}_g$ are *compatible* if $\mathcal{A}_g \cap \mathcal{B}_g = \emptyset$, $\mathcal{A}_g \subseteq \mathcal{B}_g$, or $\mathcal{B}_g \subseteq \mathcal{A}_g$.

If a clade $\mathcal{A}_g$ is on a gene tree T, then all other clades appearing on T must be compatible with $\mathcal{A}_g$.

The proof of Theorem 6 uses partitions of subsets of the taxon set $\mathcal{X}$ that occur as follows: Consider an internal node v of σ, and let $\mathcal{A} = \mathrm{desc}_{\mathcal{X}}(v)$. Then in a realization of the coalescent process on σ, some of the lineages of genes in $\mathcal{A}_g$ may coalesce below v, so that there are $|\mathcal{A}|$ or fewer lineages at v. Each such lineage determines a subset of $\mathcal{A}_g$, namely its descendants, and hence the set of lineages determines a partition of $\mathcal{A}$.

As an example, consider the species tree in Fig. 1. For the set $\mathcal{A} = \{a,b,c\}$, the partition at MRCA$(\mathcal{A})$ in both subfigures is $\{\{a\},\{b\},\{c\}\}$. Note that the partition of such a set $\mathcal{A}$ is not affected by any coalescent events occurring in the MRCA population, but only by those below. The only other partition of $\mathcal{A}$ possible for this species tree is $\{\{a,b\},\{c\}\}$. For the set $\mathcal{A} = \{a,b,c,d\}$, the partition at MRCA$(\mathcal{A})$ in Fig. 1a is $\{\{a\},\{b,c\},\{d\}\}$, and in Fig. 1b is $\{\{a,b,c\},\{d\}\}$.

Proof of Theorem 6. Suppose $\mathcal{A}$ is a clade on ψ, with $v = \mathrm{MRCA}(\mathcal{A})$. Letting $\pi(\mathcal{A}) = \{\mathcal{A}_1,\ldots,\mathcal{A}_k\}$ denote a partition of $\mathcal{A}$, we also use $\pi(\mathcal{A})$ to denote the event that the coalescent process on σ produces lineages at v defining this partition.

We will condition on this event: Specifically, recalling the notion of a coalescent history (Degnan and Salter, 2005),

$$\mathbb{P}_\sigma(\pi(\mathcal{A})) = \sum_T \sum_{\substack{\text{history } h_T, \\ h_T \text{ consistent} \\ \text{with } \pi(\mathcal{A})}} \mathbb{P}_\sigma(T,h_T). \tag{4}$$

For $\mathcal{B} \subset \mathcal{X}$ the joint probability $\mathbb{P}_\sigma(\mathcal{B}_g,\pi(\mathcal{A}))$ is computed similarly, by restricting the outer sum on the right side of Eq. (4) to those gene trees that have clade $\mathcal{B}_g$. Then

$$\mathbb{P}_\sigma(\mathcal{B}_g | \pi(\mathcal{A})) = \frac{\mathbb{P}_\sigma(\mathcal{B}_g,\pi(\mathcal{A}))}{\mathbb{P}_\sigma(\pi(\mathcal{A}))},$$

and by the law of total probability, we have the clade probability

$$\mathbb{P}_\sigma(\mathcal{B}_g) = \sum_{\pi(\mathcal{A})} \mathbb{P}_\sigma(\mathcal{B}_g \,|\, \pi(\mathcal{A}))\mathbb{P}_\sigma(\pi(\mathcal{A})).$$

Thus, to establish Eq. (3), it is enough to show that

$$\left(\sum_{\mathcal{S}\subseteq\mathcal{A}'} \mathbb{P}_\sigma(\mathcal{S}_g \cup \{A\} \cup \mathcal{C}_g \,|\, \pi(\mathcal{A}))\right) - \left(\sum_{\mathcal{S}\subseteq\mathcal{A}'} \mathbb{P}_\sigma(\mathcal{S}_g \cup \{B\} \cup \mathcal{C}_g \,|\, \pi(\mathcal{A}))\right) = 0 \tag{5}$$

holds for all choices of partition $\pi(\mathcal{A})$.

To establish Eq. (5), we show that non-zero terms cancel pairwise. However, which terms cancel depends on the partition, so for the remainder of the argument we fix $\pi(\mathcal{A})$, and assume the partition sets are indexed so that $a \in \mathcal{A}_1$.

Note first that if $b \in \mathcal{A}_1$ as well, then we are conditioning on an event that requires that the A and B lineages have coalesced into one below v. Thus, any clade on a gene tree that includes A and $\mathcal{C}_g$ must include B, because we have assumed that $\mathcal{C}$ is non-empty. Similarly, any clade that includes B and $\mathcal{C}_g$ must include A. Therefore, all probabilities in Eq. (5) are zero, so the equation holds.

Otherwise, assume $b \in \mathcal{A}_2$. We wish to give a bijective correspondence between non-zero clade probabilities in the first sum in Eq. (5) and equal clade probabilities in the second sum, with the correspondence dependent on the partition $\pi(\mathcal{A})$. That is, we wish to show that for each $\mathcal{S}_1 \subset \mathcal{A}'$, there is a corresponding $\mathcal{S}_2 \subset \mathcal{A}'$ such that

$$\mathbb{P}_\sigma[(\mathcal{S}_1)_g \cup \{A\} \cup \mathcal{C}_g \,|\, \pi(\mathcal{A})] = \mathbb{P}_\sigma[(\mathcal{S}_2)_g \cup \{B\} \cup \mathcal{C}_g \,|\, \pi(\mathcal{A})]. \tag{6}$$

Consider first the case when $(\mathcal{S}_1)_g \cup \{A\} \cup \mathcal{C}_g$ is compatible with the clades $(\mathcal{A}_1)_g, \ldots, (\mathcal{A}_k)_g$. Because $\mathcal{C}$ is non-empty, this occurs exactly when $\mathcal{S}_1 \cup \{a\}$ is the union of some of the $\mathcal{A}_i$. Thus we have

$$\mathcal{S}_1 \cup \{a\} = \mathcal{A}_1 \sqcup \bigsqcup_j \mathcal{A}_{i_j},$$

for some i_j, with all unions here disjoint. Moreover, since $b \notin \mathcal{S}_1 \cup \{a\}$, $\mathcal{A}_2$ does not appear in this expression. We therefore define $\mathcal{S}_2$ by the expression of disjoint unions

$$\mathcal{S}_2 \cup \{b\} = \mathcal{A}_2 \sqcup \bigsqcup_j \mathcal{A}_{i_j}.$$

Eq. (6) then holds, since for the coalescent process on σ above v the lineages corresponding to $\mathcal{A}_1$ and $\mathcal{A}_2$ are exchangeable. This gives us a bijection between $\mathcal{S}_1 \subset \mathcal{A}'$ and $\mathcal{S}_2 \subset \mathcal{A}'$ for which either (and hence both) of the probabilities in Eq. (6) are non-zero.

For all other $\mathcal{S}_1$, $\mathcal{S}_2$, the sets $\mathcal{S}_1 \cup \{A\} \cup \mathcal{C}_g$ and $\mathcal{S}_2 \cup \{B\} \cup \mathcal{C}_g$ are not compatible with $\pi(\mathcal{A})$, and hence these probabilities are zero. □

As a simple corollary, we immediately obtain what we call 'cherry-swapping' invariants, which express that the probability of any clade containing exactly one taxon of a 2-clade on the species tree is unchanged when that taxon is swapped out for the other taxon in the 2-clade.

Corollary 7 (*Cherry-swapping invariants*). *Suppose $\{a,b\}$ is a 2-clade on a species tree with taxa $\mathcal{X}$. Then for any $\mathcal{C} \subseteq \mathcal{X}\setminus\{a,b\}$,*

$$\mathbb{P}_\sigma(\{A\} \cup \mathcal{C}_g) - \mathbb{P}_\sigma(\{B\} \cup \mathcal{C}_g) = 0.$$

To illustrate Theorem 6, we consider next all species tree topologies on 5 or fewer taxa, and discuss invariants produced by this construction. For notational ease, we denote gene tree clades by juxtaposition of labels, rather than by sets, so, for instance $\{A,B,D,E\}$ will be denoted $ABDE$. Our focus is on those invariants associated to 3- and 4-clades, and we do not explicitly list cherry-swapping invariants except for the 3-taxon tree.

Example. For the species tree topology $\psi = ((a,b),c)$, the cherry-swapping invariant,

$$\mathbb{P}_\sigma(AC) - \mathbb{P}_\sigma(BC) = 0$$

is the only one produced by Theorem 6.

Example. For the 4-taxon caterpillar tree topology $\psi = (((a,b),c),d)$, in addition to the three cherry-swapping invariants, we find for $\mathcal{A} = \{a,b,c\}$ the invariants

$$(\mathbb{P}_\sigma(AD) + \mathbb{P}_\sigma(ABD)) - (\mathbb{P}_\sigma(CD) + \mathbb{P}_\sigma(BCD)) = 0,$$

$$(\mathbb{P}_\sigma(BD) + \mathbb{P}_\sigma(ABD)) - (\mathbb{P}_\sigma(CD) + \mathbb{P}_\sigma(ACD)) = 0,$$

$$(\mathbb{P}_\sigma(AD) + \mathbb{P}_\sigma(ACD)) - (\mathbb{P}_\sigma(BD) + \mathbb{P}_\sigma(BCD)) = 0,$$

for $\mathcal{A}' = \{b\}$, $\mathcal{A}' = \{a\}$, and $\mathcal{A}' = \{c\}$, respectively. We note that there are relations between these: the second invariant is obtained from the first by a cherry-swapping move, and the third is the sum of two cherry-swapping invariants.

For the 4-taxon balanced tree topology, $\psi = ((a,b),(c,d))$, only the six cherry-swapping invariants are obtained.

Example. If ψ is either the 5-taxon caterpillar tree topology $((((a,b),c),d),e)$, or the balanced tree topology $(((a,b),c),(d,e))$, consider $\mathcal{A} = \{a,b,c\}$.

Then for $\mathcal{A}' = \{b\}$, we obtain for various choices of $\mathcal{C}$,

$$(\mathbb{P}_\sigma(AD) + \mathbb{P}_\sigma(ABD)) - (\mathbb{P}_\sigma(CD) + \mathbb{P}_\sigma(BCD)) = 0, \tag{7}$$

$$(\mathbb{P}_\sigma(AE) + \mathbb{P}_\sigma(ABE)) - (\mathbb{P}_\sigma(CE) + \mathbb{P}_\sigma(BCE)) = 0, \tag{8}$$

$$(\mathbb{P}_\sigma(ADE) + \mathbb{P}_\sigma(ABDE)) - (\mathbb{P}_\sigma(CDE) + \mathbb{P}_\sigma(BCDE)) = 0. \tag{9}$$

Note that for the balanced species tree, Eq. (8) follows from Eq. (7) by cherry swapping D and E. However, for the caterpillar species tree, Eqs. (7) and (8) are not related by a cherry swap.

For $\mathcal{A}' = \{a\}$ we obtain Eqs. (7)–(9) again, up to cherry-swapping lineages A and B.

For $\mathcal{A}' = \{c\}$ we obtain invariants such as

$$(\mathbb{P}_\sigma(AD) + \mathbb{P}_\sigma(ACD)) - (\mathbb{P}_\sigma(BD) + \mathbb{P}_\sigma(BCD)) = 0, \tag{10}$$

but Eq. (10) is simply the sum of two cherry-swapping invariants for the cherry $\{a,b\}$, with $\mathcal{C} = \{d\}$ and $\{c,d\}$. In general, if the taxa in $\mathcal{A}\setminus\mathcal{A}'$ span a smaller clade than $\mathcal{A}$, the invariant produced will be a sum of invariants for the smaller clade. Indeed, this phenomenon occurred above, for the 4-taxon caterpillar.

Example. For the 5-taxon caterpillar topology $\psi = ((((a,b),c),d),e)$, taking $\mathcal{A} = \{a,b,c,d\}$ and using $\mathcal{A}' = \{b,c\}$ and $\mathcal{A}' = \{a,b\}$, we obtain two invariants:

$$\begin{aligned}&(\mathbb{P}_\sigma(AE) + \mathbb{P}_\sigma(ABE) + \mathbb{P}_\sigma(ACE) + \mathbb{P}_\sigma(ABCE))\\ &\quad -(\mathbb{P}_\sigma(DE) + \mathbb{P}_\sigma(BDE) + \mathbb{P}_\sigma(CDE) + \mathbb{P}_\sigma(BCDE)) = 0\end{aligned}$$

and

$$\begin{aligned}&(\mathbb{P}_\sigma(CE) + \mathbb{P}_\sigma(ACE) + \mathbb{P}_\sigma(BCE) + \mathbb{P}_\sigma(ABCE))\\ &\quad -(\mathbb{P}_\sigma(DE) + \mathbb{P}_\sigma(ADE) + \mathbb{P}_\sigma(BDE) + \mathbb{P}_\sigma(ABDE)) = 0.\end{aligned}$$

Other choices of $\mathcal{A}'$ give only invariants in the space spanned by those previously discussed.

Example. For the 5-taxon pseudo-caterpillar tree topology $\psi = (((a,b),(d,e)),c)$, taking $\mathcal{A} = \{a,b,d,e\}$ we obtain

$$\begin{aligned}&(\mathbb{P}_\sigma(AC) + \mathbb{P}_\sigma(ABC) + \mathbb{P}_\sigma(ACE) + \mathbb{P}_\sigma(ABCE))\\ &\quad -(\mathbb{P}_\sigma(CD) + \mathbb{P}_\sigma(BCD) + \mathbb{P}_\sigma(CDE) + \mathbb{P}_\sigma(BCDE)) = 0,\end{aligned} \tag{11}$$

and three other invariants that can also be obtained by cherry swapping from Eq. (11). Since $\mathbb{P}_\sigma(ACE) = \mathbb{P}_\sigma(BCD)$ by cherry swapping, two of the eight terms can be cancelled.

Remark. All the linear invariants above for 3-, 4-, and 5-taxon trees are, of course, among those that can be found computationally. Gröbner basis calculations do not necessarily produce exactly these, but by cherry-swapping and taking suitable linear combinations of computed linear invariants, all of these appear. However, at least for trees on four and five taxa, there are additional linear invariants beyond the ones of Theorem 6. We give these in Appendix B, as it would be interesting to have non-computational means of obtaining them, as well as the higher degree invariants.

6. Identifying clades

Suppose we are given the clade probabilities $\{\mathbb{P}_\sigma(\mathcal{C}_g)\}$ arising from the multispecies coalescent on an unknown species tree σ, and we wish to know if σ displays a particular clade. By the results of Section 4, high probability may identify some clades on σ. However, it remains to be seen how one might identify clades on σ that have lower probability of occurring on gene trees as a result of high levels of incomplete lineage sorting.

From Section 5 we know that if $\mathcal{A}$ is a clade on σ then for every non-empty subset $\mathcal{C} \subseteq \mathcal{X} \backslash \mathcal{A}$, and every $a,b \in \mathcal{A}$, the linear invariant associated to $\mathcal{A}$, $\mathcal{C}$, a, and b vanishes. For these invariants to be useful for identifying clades, however, we must also know that if σ does not display the clade $\mathcal{A}$, then one of these invariants does not vanish.

Lemma 8. *Suppose $\mathcal{A}$ is a non-trivial clade on a species tree σ, and $a \in \mathcal{A}$ and $b \in \mathcal{X} \backslash \mathcal{A}$. Let $\mathcal{B}$ denote the set obtained by replacing a with b in $\mathcal{A}$, that is, $\mathcal{B} = (\mathcal{A} \backslash \{a\}) \cup \{b\}$. Then $\mathbb{P}_\sigma(\mathcal{A}_g) > \mathbb{P}_\sigma(\mathcal{B}_g)$.*

Proof. Let $v = \mathrm{MRCA}(\mathcal{A} \cup \{b\})$ on σ. Let $\mathcal{A}' = \mathcal{A} \backslash \{a\}$, so $\mathcal{A} = \{a\} \cup \mathcal{A}'$ and $\mathcal{B} = \{b\} \cup \mathcal{A}'$.

Then, using the phrase 'X coalesces with $\mathcal{Y}$ above v' to mean the lineage of X first coalesces with any gene lineage in a set $\mathcal{Y}$ in a population in the species tree above the node v,

$$\mathbb{P}_\sigma(\mathcal{A}_g) \tag{12}$$

$$= \mathbb{P}_\sigma(\{A\} \cup \mathcal{A}'_g) \tag{13}$$

$$= \mathbb{P}_\sigma(\{A\} \cup \mathcal{A}'_g \text{ and } A \text{ coalesces with } \mathcal{A}'_g \text{ below } v) \tag{14}$$

$$+ \mathbb{P}_\sigma(\{A\} \cup \mathcal{A}'_g \text{ and } A \text{ coalesces with } \mathcal{A}'_g \text{ above } v)$$

$$> \mathbb{P}_\sigma(\{A\} \cup \mathcal{A}'_g \text{ and } A \text{ coalesces with } \mathcal{A}'_g \text{ above } v) \tag{15}$$

$$\geq \mathbb{P}_\sigma(\{A\} \cup \mathcal{A}'_g,\ A \text{ coalesces with } \mathcal{A}'_g \text{ above } v, \tag{16}$$

$$\text{and } B \text{ coalesces with } \mathcal{X}_g \backslash \{B\} \text{ above } v)$$

$$= \mathbb{P}_\sigma(\{B\} \cup \mathcal{A}'_g,\ A \text{ coalesces with } \mathcal{A}'_g \text{ above } v, \tag{17}$$

$$\text{and } B \text{ coalesces with } \mathcal{X}_g \backslash \{B\} \text{ above } v)$$

$$= \mathbb{P}_\sigma(\{B\} \cup \mathcal{A}'_g) \tag{18}$$

$$= \mathbb{P}_\sigma(\mathcal{B}_g). \tag{19}$$

Lines (16) and (17) are equal due to exchangeability of lineages; given any sequence of coalescences in the event of line (16), there is an equally probable sequence of coalescences in the event of line (17) in which B coalesces in A's place to form $\{B\} \cup \mathcal{A}'_g$ instead of $\{A\} \cup \mathcal{A}'_g$. Lines (17) and (18) are equal because the event that $\{B\} \cup \mathcal{A}'_g$ is a clade can only occur when A and B each coalesce as described in (17), and thus these extra statements are redundant.

Thus, $\mathbb{P}_\sigma(\mathcal{A}_g) > \mathbb{P}_\sigma(\mathcal{B}_g)$. □

Remark. One might wish to extend the above result to sets obtained by replacing k elements in a clade with k elements outside it. However, simple examples show that this is impossible. For instance, if $\sigma = (((a,b):x,c):y,(d,e):z)$ where z is large and both x and y are small, then one can have $\mathbb{P}_\sigma(ABC) < \mathbb{P}_\sigma(CDE)$. For example, if $(x,y,z) = (0.05, 0.05, 2.0)$, then the highest probability clades are DE, AB, AC, BC, CDE, and ABC with probabilities 0.889, 0.269, 0.220, 0.220, 0.194, and 0.188, respectively (computed by COAL, Degnan and Salter, 2005). Thus for these branch lengths, we have $\mathbb{P}_\sigma(ABC) < \mathbb{P}_\sigma(CDE)$, and the greedy strategy of accepting the most probable clades one-at-a-time returns the non-matching tree $((a,b),(c,(d,e)))$.

The same example shows that for a set $\mathcal{C} \subset \mathcal{X}$ there can exist $y \in \mathcal{C}$ such that for all $x \in \mathcal{X} \setminus \mathcal{C}$, $\mathbb{P}_\sigma(\mathcal{C}_g) > \mathbb{P}_\sigma((\mathcal{C} \backslash \{y\}) \cup \{x\})$ and yet $\mathcal{C}$ is not a clade. In this example, $\{c,d,e\}$ is not a clade on the species tree, yet CDE is more probable than ADE or BDE on gene trees.

Lemma 8 allows us to show that if all cherry-swapping invariants are satisfied for a particular candidate 2-clade, it is in fact a 2-clade on the species tree.

Proposition 9 (*Clade probabilities determine species 2-clades*). *For $\sigma = (\psi, \lambda)$ an n-taxon binary species tree on a set of taxa $\mathcal{X}$, the 2-clades of ψ are identifiable from clade probabilities. In particular, for any $a,b \in \mathcal{X}$, $\{a,b\}$ is a clade on ψ if, and only if, for every $\mathcal{D} \subseteq \mathcal{X} \backslash \{a,b\}$, $\mathbb{P}_\sigma(\{A\} \cup \mathcal{D}_g) = \mathbb{P}_\sigma(\{B\} \cup \mathcal{D}_g)$.*

Proof. If $\{a,b\}$ is a clade on σ, then by Corollary 7, $\mathbb{P}(\{A\} \cup \mathcal{D}_g) = \mathbb{P}(\{B\} \cup \mathcal{D}_g)$ for any taxon set $\mathcal{D}$ not containing a or b.

Suppose now that $\{a,b\}$ is not a clade on ψ. Then, because σ is binary, at least one of a or b (let us say a) is in a non-trivial clade $\mathcal{C}$ on ψ that excludes the other. Let $\mathcal{D} = \mathcal{C} \backslash \{a\}$.

By Lemma 8, $\mathbb{P}_\sigma(\{A\} \cup \mathcal{D}_g) \neq \mathbb{P}_\sigma(\{B\} \cup \mathcal{D}_g)$. □

For clades of more than two taxa on a species trees, we obtain a slightly weaker result: As long as the edge length vector λ does not lie in a set of measure zero, then the clades on the species tree can be identified. The first step toward this result is the following.

Lemma 10. *Let ψ be a species tree topology on $\mathcal{X}$, and $\mathcal{X} = \mathcal{A} \sqcup \mathcal{D}$ a disjoint union of non-empty subsets with $|\mathcal{A}| \geq 2$. Then if $\mathcal{A}$ is not a clade on ψ and* MRCA($\mathcal{A}$) *is a binary node, then there exists some $\mathcal{C} \subseteq \mathcal{D}$, $a,b \in \mathcal{A}$, and some choice of edge lengths λ such that the corresponding clade invariant of Theorem 6 does not vanish on the clade probabilities arising under the multispecies coalescent on $\sigma = (\psi, \lambda)$.*

Proof. Suppose $\mathcal{A}$ is not a clade on ψ. Let $v = \mathrm{MRCA}(\mathcal{A})$ on ψ, so $\mathcal{E} = \mathrm{desc}_{\mathcal{X}}(v) \backslash \mathcal{A}$ is non-empty. One or both children nodes w_1, w_2 of v have an element of $\mathcal{E}$ as a descendant, so we may assume $\mathcal{C} = \mathrm{desc}_{\mathcal{X}}(w_1) \cap \mathcal{E}$ is non-empty. Let $a \in \mathcal{A} \cap \mathrm{desc}_{\mathcal{X}}(w_1)$, $b \in \mathcal{A} \cap \mathrm{desc}_{\mathcal{X}}(w_2)$. Consider the clade invariant of Theorem 6 associated to $\mathcal{A}, \mathcal{C}, a, b$.

We next give edge lengths λ for which this invariant will not vanish at the clade probabilities arising from the multispecies coalescent on $\sigma = (\psi, \lambda)$. Let all internal edges of ψ below v have length (near) 0 except the edge (v, w_1) which is assigned length (near) ∞. Lengths of edges above v can be fixed at any finite nonzero values.

With these assignments, the only partition of $\mathrm{desc}_{\mathcal{X}}(v)$ according to lineages at v that appears with non-negligible probability is that with $\mathrm{desc}_{\mathcal{X}}(w_1)$ forming one partition set, while all elements of $\mathrm{desc}_{\mathcal{X}}(w_2)$ are in singleton sets. But since $\mathcal{C} \subset \mathrm{desc}_{\mathcal{X}}(w_1)$, $a \in \mathrm{desc}_{\mathcal{X}}(w_1)$ and b is in a singleton set, the only clades that can result with non-negligible probability that contain both B and elements of $\mathcal{C}_g$ must also contain A. Thus all the clades appearing in the second term of Eq. (3) have probability arbitrarily close to 0. However, the clade $(\mathrm{desc}_{\mathcal{X}}(w_1))_g$ appears in the first term and has non-negligible probability. Thus Eq. (3) is violated. □

Theorem 11. *Let ψ be a rooted binary species tree topology on $\mathcal{X}$, where $\mathcal{X}=\mathcal{A}\sqcup\mathcal{D}$ is a disjoint union of non-empty subsets. If $\mathcal{A}$ is not a clade on ψ then for all choices of edge lengths λ except those in some set of measure zero there exists some $\mathcal{C}\subseteq\mathcal{D}$, $a,b\in\mathcal{A}$, such that the corresponding clade invariant does not vanish on the clade probabilities arising under the multispecies coalescent on $\sigma=(\psi,\lambda)$.*

Proof. The clade probabilities arising from σ can be expressed as polynomials in the exponentials of the negatives of the interior edge lengths. By Lemma 10, there is an invariant which, when composed with this polynomial map, does not vanish at some point in the space $(0,1]^{n-2}$ of these exponentials. But since this composition is a polynomial, its non-vanishing at some point implies the set where it vanishes has measure zero in $(0,1]^{n-2}$. Mapping this set to interior edge lengths by $-\log(x)$ shows the set of edge lengths for which the invariant vanishes has measure zero. □

Since, except for a negligible set of edge length parameters, whether a species tree has a particular clade can be tested by examining clade probabilities, one can similarly determine the full species tree topology.

Though intriguing, we have not investigated whether the set of measure zero in Theorem 11 is non-empty, and thus whether there exist (rare) instances in which the vanishing of clade invariants might erroneously lead one into suspecting a clade occurs on the species tree.

Corollary 12. *Let ψ be a rooted binary species tree topology on $\mathcal{X}$. For generic choices of edge lengths λ, ψ can be identified from the probabilities of clades under the multispecies coalescent on $\sigma=(\psi,\lambda)$.*

Proof. For any subset of taxa $\mathcal{A}\subset\mathcal{X}$, if we find any invariant given by Theorem 6 that fails to vanish on the clade probabilities for $\sigma=(\psi,\lambda)$, then $\mathcal{A}$ is not a clade on ψ. If all such invariants vanish, then by Theorem 11, either $\mathcal{A}$ is a clade on ψ, or λ lies in a set of measure zero (which is dependent on $\mathcal{A}$,$\mathcal{C}$,a, and b used in defining the invariant).

Thus, considering all proper subsets $\mathcal{A}$ of $\mathcal{X}$, we can determine all clades, unless the edge lengths λ lie in a set of measure zero (the finite union of sets of measure zero for each invariant).

Finally, the clades of ψ determine ψ. □

Remark. If one considers a non-binary species tree to be specified by the topology of an arbitrarily chosen binary resolution of it, along with the assignment of edge length 0 to any introduced edges, then both Theorem 11 and Corollary 12 still apply. Indeed, the special choices of some 0 edge lengths form a set of Lebesgue measure zero in the full set of possible edge lengths, so regardless of whether such trees can be identified, the statements remain valid.

A particular feature of non-binary species trees that is identifiable is a *k-cherry*, a set of $k\geq 2$ leaves $\{x_1\ldots,x_k\}\in\mathcal{X}$ that all share a common parent node and form a clade. This will prove useful for identifying the extended species trees defined in Section 2, which describes the sampling of multiple individuals per taxon.

Proposition 13 (*Clade probabilities determine extended species tree k-cherries*). *Let $\sigma^*=(\psi^*,\lambda,\delta)$ be an extended species tree on $\mathcal{X}$ for which the pruned species tree ψ is binary. Then the k-cherries of ψ^* are identifiable from gene clade probabilities from the multispecies coalescent on σ^* for all choices of edge lengths λ outside a set of measure zero.*

In particular, $\{x_{i_1j_1},\ldots,x_{i_kj_k}\}\subseteq\mathcal{X}^*$ is a k-cherry on ψ^* if, and only if, it is a maximal subset of $\mathcal{X}^*$ such that for every $1\leq l<m\leq k$ and every $y\in\mathcal{X}^*\backslash\{x_{i_lj_l},x_{i_mj_m}\}$,

$$\mathbb{P}_\sigma(\{X_{i_lj_l},Y\})=\mathbb{P}_\sigma(\{X_{i_mj_m},Y\}).$$

Proof. Let $\mathcal{K}=\{x_{i_1j_1},\ldots,x_{i_kj_k}\}$ be a k-cherry on ψ^*, with MRCA the node v. Then for any $y\in\mathcal{X}^*\backslash\{x_{i_lj_l},x_{i_mj_m}\}$, $\mathbb{P}_\sigma(\{X_{i_lj_l},Y\})=\mathbb{P}_\sigma(\{X_{i_mj_m},Y\})$ by the exchangeability of $X_{i_lj_l}$ and $X_{i_mj_m}$.

To see that $\mathcal{K}$ is maximal with respect to this property, suppose $z\in\mathcal{X}^*\backslash\mathcal{K}$. (If no such z exists, maximality is clear.) We show $\mathcal{K}$ cannot be augmented by z by showing that $\mathbb{P}_\sigma(\{X_{i_1j_1},X_{i_2j_2}\})\neq\mathbb{P}_\sigma(\{Z,X_{i_2j_2}\})$ for some choice of λ. This then implies the same statement for generic values of λ, since these probabilities are polynomials in the exponentials of negative branch lengths.

Choose all internal branch lengths of the species tree to be (near) 0 except for the branch e above v, which we choose to have length (near) ∞. Consider the event E that the $X_{i_1j_1}$ and $X_{i_2j_2}$ lineages coalesce on e and are the first of the $\mathcal{K}$ lineages to do so. Then one sees that

$$\mathbb{P}_\sigma(\{X_{i_1j_i},X_{i_2j_2}\})>\mathbb{P}_\sigma(E)\approx\binom{k}{2}^{-1},$$

where the approximation becomes increasingly accurate as more extreme branch lengths are chosen. However such choices of branch lengths make $\mathbb{P}_\sigma(\{Z,X_{i_2j_2}\})$ as close to 0 as desired, since the probability of the clade $\mathcal{K}$ goes to 1, and this is incompatible with clade $\{Z,X_{i_2j_2}\}$. Thus $\mathbb{P}_\sigma(\{X_{i_1j_1},X_{i_2j_2}\})\neq\mathbb{P}_\sigma(\{Z,X_{i_2j_2}\})$.

To establish the converse, suppose now that $\mathcal{K}$ is maximal with respect to the stated property, but is not a k-cherry. By the above argument, maximality implies $\mathcal{K}$ is not a subset of any l-cherry for $l>k$.

To achieve a contradiction, it is sufficient to show that there exist $x_{i_1j_1},x_{i_2j_2}\in\mathcal{K}$, $y\in\mathcal{X}^*$ such that $\mathbb{P}_\sigma(\{X_{i_1j_1},Y\})\neq\mathbb{P}_\sigma(\{X_{i_2j_2},Y\})$ unless branch lengths lie on a set of measure 0. Let $v=\mathrm{MRCA}(\mathcal{K})$. Since $\mathcal{K}$ is not contained in an l-cherry, there exists a non-leaf node w which is a child of v. Moreover, v is binary, since ψ is.

Choose $x_{i_1j_1}\in\mathcal{K}\backslash\mathrm{desc}_{\mathcal{X}}(w)$, which is non-empty because $v=\mathrm{MRCA}(\mathcal{K})$. The node w has at least two distinct leaf descendants, and since v is binary at least one leaf descendant of w must be in $\mathcal{K}$. Choose $x_{i_2j_2}\in\mathcal{K}\cap\mathrm{desc}_{\mathcal{X}}(w)$, and $y\in\mathrm{desc}_{\mathcal{X}}(w)\backslash\{x_{i_2j_2}\}$.

Let $m=|\mathrm{desc}_{\mathcal{X}}(w)|$. By choosing the length of the edge connecting w and v to be near ∞, and the length of all edges descended from w to be near 0, the probability of clade $\{X_{i_1j_1},Y\}$ will be arbitrarily close to 0, and the probability of the clade $\{X_{i_2j_2},Y\}$ will be bounded below by a number arbitrarily close to $\binom{m}{2}^{-1}$, as in the argument above. Thus, there exist branch lengths with $\mathbb{P}_\sigma(\{X_{i_1j_1},Y\})\neq\mathbb{P}_\sigma(\{X_{i_2j_2},Y\})$. Because these probabilities are polynomials (in transformed branch lengths), the set of branch lengths where $\mathbb{P}_\sigma(\{X_{i_1j_1},Y\})=\mathbb{P}_\sigma(\{X_{i_2j_2},Y\})$ has measure 0. □

Finally, we apply Proposition 13 to show generic identifiability of species trees from clade probabilities when there are $\delta_i\geq 1$ lineages sampled for taxon x_i. (See Fig. 2.)

Corollary 14. *Let $\sigma^*=(\psi^*,\lambda,\delta)$ be an extended species tree, for which the pruned species tree ψ is binary. For generic choices of edge lengths λ, the topology of ψ^* can be identified from the probabilities of clades under the multispecies coalescent.*

Proof. All k-cherries of ψ^* can be identified by Proposition 13 (although this is unnecessary if one assumes species assignments are given). By assumption there are no other polytomies on ψ^*;

for any other clade $\mathcal{A}$ on the extended species tree, $v = \mathrm{MRCA}(\mathcal{A})$ is a binary node. Thus Theorem 6 and Lemma 10 apply. These imply that for generic λ all clades on the extended species tree can be identified, and hence ψ^* can be identified. □

Since this corollary does not assume that the assignment of individuals to taxa is known in advance, it implies that under some circumstances species assignment can be deduced from clade probabilities. In particular, any taxon for which three or more individuals are sampled will be identifiable from ψ^*. However taxa in which two individuals have been sampled will be indistinguishable from two taxa forming a 2-clade with one individual sampled from each.

7. Discussion

We have shown that for generic branch lengths on a binary species tree, it is possible to identify clades of the species tree, and therefore the species tree topology, from probabilities of clades on gene trees. More generally, we showed identifiability of clades consisting of taxa descended from binary nodes even if the species tree is not itself binary. In addition, we investigated how probable a clade on a gene tree must be to infer it is also a clade on the species tree.

We have not shown the identifiability of branch lengths from clade probabilities. However, for any given species tree topology it is possible to write systems of equations of clade probabilities as functions of the branch lengths. (As examples, consider the systems of equations for clade probabilities for some 4-taxon species trees shown in Table A1.) These systems are non-linear but polynomial in the transformed branch lengths. Since the number of branch length parameters is $n-2$ for an n-taxon tree and there are 2^n-n-2 non-trivial clade probabilities, it is reasonable to expect such systems to be solvable, in principle, for any sized tree. Although for particular small trees these can be solved, we have not found a general method applicable to arbitrary trees. It thus remains conjectural that species tree branch lengths are identifiable from clade probabilities. If multiple individuals are sampled from some species, then the species tree has additional branch length parameters (for branches leading to such species), but will have an even larger number of clades' probabilities that could conjecturally be used to estimate branch lengths.

While the invariants of Theorem 6 are useful in proving identifiability of a species tree topology, they do not immediately indicate a practical way to infer the species tree from clade probabilities. In particular, each term in the invariant of Theorem 6 is the probability that a random gene tree has a clade that is not a clade on the species tree. The clade probabilities needed for the invariant of Theorem 6 may therefore be quite small. For species trees with moderately long branches, many of these probabilities could be difficult to estimate from finite data sets. However, in such a situation the results of Section 4 might offer an alternative way of inferring species tree clades as those which occur with high frequency on gene trees. This suggests the possibility of a hybrid approach in which one accepts highly probable clades as being clades on the species tree, as in a greedy algorithm, yet exploits the symmetries of clade probabilities expressed by invariants to determine other species clades. Thus our identifiability results should motivate further research on species tree inference methods that are statistically consistent and that can outperform greedy consensus on typical data sets with imperfectly estimated clade probabilities.

Acknowledgment

The authors thank the Statistical and Applied Mathematical Sciences Institute, where this work was begun during its 2008–2009 program on Algebraic Methods in Systems Biology and Statistics, and the Institut Mittag-Leffler, where the writing was completed during the Spring 2011 program on Algebraic Geometry with a View Towards Applications. E.S.A. and J.A.R. were supported by funds from the National Science Foundation, Grant DMS 0714830, and J.A.R. by an Erskine Fellowship from the University of Canterbury. J.H.D. was funded by the New Zealand Marsden Fund. All authors contributed equally to this work.

Appendix A. Clade probabilities for subtrees as linear combinations

An essential difficulty in dealing with clade probabilities in mathematical arguments is that it is not easy to see relationships between probabilities of clades on gene trees arising from a species tree on a set of taxa and the clade probabilities on the induced gene trees obtained by restricting the set of taxa to a smaller set. This frustrates the common approach used to prove results for large trees, by inductive arguments on the number of taxa.

Consider a set of taxa $\mathcal{X}$, a proper subset $\mathcal{Y} \subset \mathcal{X}$, and a species tree $\sigma = (\psi,\lambda)$ on $\mathcal{X}$. We show here that, in general, probabilities of gene tree clades for the induced species tree on $\mathcal{Y}$, $\sigma|\mathcal{Y}$, cannot be written as the same linear combination of gene tree clade probabilities for σ for all choices of ψ. Thus there is no linear formula for the clade probabilities for the smaller taxon set that does not depend on the species tree topology.

This is in contrast to, for example, gene tree probabilities for $\sigma|\mathcal{Y}$, which can be written as linear combinations of gene tree probabilities for σ, where the weight assigned to each gene tree probability in the combination has no dependency on σ.

To show that the same linear combination cannot be used to marginalize clade probabilities independently of the species tree, we consider three species tree topologies on four taxa, as given in Table A1: $(((a,b),c),d)$, $((a,d),(b,c))$, and $((a,b),(c,d))$. Clade probabilities can be obtained from Degnan et al. (2009, Table D1). There are 10 non-trivial clades, so any linear combination of clade probabilities $c_1,\ldots,c_{10}$ has the form

$$\sum_{i=1}^{10} \alpha_i c_i \tag{A.1}$$

for some $\alpha_1,\ldots,\alpha_{10}$.

We consider obtaining the probability of clade CD when the taxon set is restricted to $\{a,c,d\}$ (*i.e.*, marginalizing over taxon b). Assuming first that the species tree is $((a,d):x,(b,c):y)$, the restricted species tree is $((a,d):x,c)$, and the probability of clade CD is $\frac{1}{3}X$. If a linear combination of the clade probabilities on the

Table A1
Probabilities of clades under three 4-taxon species trees. $X = \exp(-x)$, $Y = \exp(-y)$.

Clade	Probability under species tree		
	$(((a,b):x,c):y,d)$	$((a,d):x,(b,c):y)$	$((a,b):x,(c,d):y)$
$c_1 = \mathbb{P}_\sigma(AB)$	$1-\frac{2}{3}X-\frac{1}{9}XY^3$	$\frac{2}{9}XY$	$1-\frac{2}{3}X-\frac{1}{9}XY$
$c_2 = \mathbb{P}_\sigma(AC)$	$\frac{1}{3}X-\frac{1}{9}XY^3$	$\frac{2}{9}XY$	$\frac{2}{9}XY$
$c_3 = \mathbb{P}_\sigma(AD)$	$\frac{1}{6}XY+\frac{1}{18}XY^3$	$1-\frac{2}{3}X-\frac{1}{9}XY$	$\frac{2}{9}XY$
$c_4 = \mathbb{P}_\sigma(BC)$	$\frac{1}{3}X-\frac{1}{9}XY^3$	$1-\frac{2}{3}Y-\frac{1}{9}XY$	$\frac{2}{9}XY$
$c_5 = \mathbb{P}_\sigma(BD)$	$\frac{1}{6}XY+\frac{1}{18}XY^3$	$\frac{2}{9}XY$	$\frac{2}{9}XY$
$c_6 = \mathbb{P}_\sigma(CD)$	$\frac{1}{3}Y-\frac{1}{6}XY+\frac{1}{18}XY^3$	$\frac{2}{9}XY$	$1-\frac{2}{3}Y-\frac{1}{9}XY$
$c_7 = \mathbb{P}_\sigma(ABC)$	$1-\frac{2}{3}Y-\frac{1}{3}XY+\frac{1}{6}XY^3$	$\frac{1}{3}X-\frac{1}{6}XY$	$\frac{1}{3}Y-\frac{1}{6}XY$
$c_8 = \mathbb{P}_\sigma(ABD)$	$\frac{1}{3}Y-\frac{1}{6}XY$	$\frac{1}{3}Y-\frac{1}{6}XY$	$\frac{1}{3}Y-\frac{1}{6}XY$
$c_9 = \mathbb{P}_\sigma(ACD)$	$\frac{1}{6}XY$	$\frac{1}{3}Y-\frac{1}{6}XY$	$\frac{1}{3}X-\frac{1}{6}XY$
$c_{10} = \mathbb{P}_\sigma(BCD)$	$\frac{1}{6}XY$	$\frac{1}{3}X-\frac{1}{6}XY$	$\frac{1}{3}X-\frac{1}{6}XY$

larger tree is to yield this probability, then by inserting the formulas for the c_i from Table A1 into Eq. (A.1) and equating coefficients, we obtain the following equations:

$$\alpha_3+\alpha_4=0,$$

$$-2\alpha_3+\alpha_7+\alpha_{10}=1,$$

$$-2\alpha_4+\alpha_8+\alpha_9=0,$$

$$4\alpha_1+4\alpha_2-2\alpha_3-2\alpha_4+4\alpha_5+4\alpha_6-3\alpha_7-3\alpha_8-3\alpha_9-3\alpha_{10}=0,$$

where the rows correspond to the coefficients of 1, X, Y, and XY. The system is underdetermined since there are 10 unknowns and only four equations.

Similar systems can be obtained by considering other species trees. For the other species trees in Table A1, $(((a,b):x,c):y,d)$ and $((a,b):x,(c,d):y)$, respectively, restricting to taxa $\{a,c,d\}$ leads to trees $((a,c):y,d)$ and $(a,(c,d):y)$, and probabilities of clade CD that are $\frac{1}{3}Y$ and $1-\frac{2}{3}Y$. Equating coefficients on all three species trees in Table A1, we have the equations encoded by the following 13×11 augmented matrix:

$$\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0\\
-2 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 & 0 & 0 & 1\\
0 & 0 & 1 & 0 & 1 & -1 & -2 & -1 & 1 & 1 & 0\\
-2 & -2 & 1 & -2 & 1 & 1 & 3 & 0 & 0 & 0 & 0\\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & -2 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1\\
0 & 0 & 0 & -2 & 0 & 0 & 0 & 1 & 1 & 0 & 0\\
4 & 4 & -2 & -2 & 4 & 4 & -3 & -3 & -3 & -3 & 0\\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1\\
-2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0\\
0 & 0 & 0 & 0 & 0 & -2 & 1 & 1 & 0 & 0 & -2\\
-2 & 4 & 4 & 4 & 4 & -2 & -3 & -3 & -3 & -3 & 0
\end{bmatrix}.$$

Here rows 1–5 represent the system of equations implied by the species tree $(((a,b),c),d)$, rows 6–9 represent the system of equations corresponding to the species tree $((a,d),(b,c))$, and rows 10–13 represent the system of equations corresponding to the species tree $((a,b),(c,d))$. Gaussian elimination shows this system of 13 equations is inconsistent.

Appendix B. Additional clade invariants for small trees

For trees on five or fewer taxa, computations of a Gröbner basis for invariants in clade probabilities show that the construction of Theorem 6 fails to produce all invariants, or even all linear ones. In this appendix, we indicate the results of such computations that we performed using the software `Singular` (Greuel et al., 2009). We emphasize that by linear invariant we mean linear homogeneous invariant, so that the trivial invariant, which is inhomogeneous, is not counted when we give dimensions of spaces.

For the 3-taxon tree there is only a single invariant, the linear one arising from cherry-swapping, produced by Theorem 6.

For the 4-taxon balanced tree topology $((a,b),(c,d))$, there is a six-dimensional space of linear invariants, yet the ones constructed in Theorem 6 span only a five-dimensional subspace. The additional generator needed to obtain all linear invariants can be taken to be

$$\mathbb{P}_\sigma(AB)-\mathbb{P}_\sigma(CD)-2\mathbb{P}_\sigma(ABC)+2\mathbb{P}_\sigma(ACD).$$

The ideal of all invariants has just one additional generator, which is quadratic.

For the 4-taxon caterpillar tree topology $(((a,b),c),d)$, there is a five-dimensional space of linear invariants. However the construction of Theorem 6 produces only a four-dimensional space of linear invariants. For the full space of linear invariants, the polynomial

$$\mathbb{P}_\sigma(AB)+2\mathbb{P}_\sigma(AC)+9\mathbb{P}_\sigma(CD)-\mathbb{P}_\sigma(ABC)\\ -11\mathbb{P}_\sigma(ABD)-4\mathbb{P}_\sigma(ACD)$$

can be taken as the missing generator.

In addition, there were one quadratic and three cubic polynomials in a full Gröbner basis.

For the 5-taxon balanced tree topology $(((a,b),c),(d,e))$, the construction of Theorem 6 produces a 14-dimensional subspace within a 16-dimensional space of linear invariants. Additional generators can be taken to be

$$22\mathbb{P}_\sigma(CD)+5\mathbb{P}_\sigma(DE)-5\mathbb{P}_\sigma(ABC)-22\mathbb{P}_\sigma(ABD)\\ +15\mathbb{P}_\sigma(CDE)+10\mathbb{P}_\sigma(ABCD)-25\mathbb{P}_\sigma(ABDE)-20\mathbb{P}_\sigma(ACDE)$$

and

$$11\mathbb{P}_\sigma(AB)+22\mathbb{P}_\sigma(AC)-25\mathbb{P}_\sigma(DE)+14\mathbb{P}_\sigma(ABC)\\ -22\mathbb{P}_\sigma(ABD)-44\mathbb{P}_\sigma(ACD)+24\mathbb{P}_\sigma(CDE)-50\mathbb{P}_\sigma(ABCD)\\ +4\mathbb{P}_\sigma(ABDE)+56\mathbb{P}_\sigma(ACDE).$$

In addition to the linear invariants, there are eight quadratic invariants and 13 cubic invariants in a Gröbner basis for the ideal.

For the 5-taxon pseudo-caterpillar tree topology $(((a,b),(d,e)),c)$, the construction of Theorem 6 produces a 13-dimensional subspace within a 14-dimensional space of linear invariants. An additional generator can be taken to be

$$\mathbb{P}_\sigma(AB)-\mathbb{P}_\sigma(DE)-6\mathbb{P}_\sigma(ABC)-2\mathbb{P}_\sigma(ABD)+2\mathbb{P}_\sigma(ADE)+6\mathbb{P}_\sigma(CDE).$$

The algorithm for computing the full ideal of invariants for this topology did not terminate in a reasonable amount of time, so the full ideal remains unknown. Partial computations in which the degree of generators is bounded show that there are generators in degrees 2, 3, 4, 5, and 6, in addition to linear invariants.

For the 5-taxon caterpillar tree $((((a,b),c),d),e)$, the construction above produces a 11-dimensional subspace within a 12-dimensional space of linear invariants. One choice for the additional generator is

$$5\mathbb{P}_\sigma(AB)+10\mathbb{P}_\sigma(AC)+24\mathbb{P}_\sigma(CD)+62\mathbb{P}_\sigma(DE)\\ +2\mathbb{P}_\sigma(ABC)-20\mathbb{P}_\sigma(ABD)-29\mathbb{P}_\sigma(ABE)+8\mathbb{P}_\sigma(ACD)\\ -58\mathbb{P}_\sigma(ACE)+45\mathbb{P}_\sigma(CDE)-7\mathbb{P}_\sigma(ABCD)\\ -76\mathbb{P}_\sigma(ABCE)-44\mathbb{P}_\sigma(ABDE)+2\mathbb{P}_\sigma(ACDE).$$

Our attempt to compute a Gröbner basis for the caterpillar topology did not terminate in a reasonable amount of time. We did, however, find quadratic generators in addition to the linear ones, but found no higher degree generators. It is reasonable to speculate that the full ideal is generated in degree one and two for this topology.

It would be quite interesting to find general constructions that lead to the additional linear invariants not explained by Theorem 6. Similarly, understanding the structure of higher degree invariants by non-computational means is an open challenge.

References

Allman, E.S., Degnan, J.H., Rhodes, J.A., 2011. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J. Math. Biol. 62 (6), 833–862.

Ané, C., Larget, B., Baum, D.A., Smith, S.D., Rokas, A., 2007. Bayesian estimation of concordance factors. Mol. Biol. Evol. 24, 412–426.

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. DIMACS. AMS, pp. 163–183.

Cranston, K.A., Hurwitz, B., Ware, D., Stein, L., Wing, R.A., 2009. Species trees from highly incongruent gene trees in rice. Syst. Biol. 58, 489–500.

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

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

Degnan, J.H., Salter, L.A., 2005. Gene tree distributions under the coalescent process. Evolution 59, 24–37.

Ebersberger, I., Galgoczy, P., Taudien, S., Taenzer, S., 2007. Mapping human genetic ancestry. Mol. Biol. Evol. 24, 2266–2277.

Edwards, S.V., 2009. Is a new and general theory of systematics emerging? Evolution 63, 1–19.

Eldon, B., Wakeley, J., 2006. Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics 172, 2621–2633.

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

Greuel, G.-M., Pfister, G., Schönemann, H., 2009. Singular 3.1.0—A Computer Algebra System for Polynomial Computations. Technical Report, Centre for Computer Algebra, University of Kaiserslautern ⟨http://www.singular.uni-kl.de⟩.

Knowles, L.L., Kubatko, L.S. (Eds.), 2010. Estimating Species Trees: Practical and Theoretical Aspects. Wiley-Blackwell, Hoboken, NJ.

Larget, B.R., Kotha, S.K., Dewey, C.N., Ané, C., 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26, 2910–2911.

Liu, L., Yu, L., Edwards, S.V., 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10, 302.

Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, NY.

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

Pitman, J., 1999. Coalescents with multiple collisions. Ann. Prob. 27, 1870–1902.

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

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.

Rosenberg, N.A., 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61, 225–247.

Semple, C., Steel, M., 2003. Phylogenetics. Oxford University Press, Oxford, UK.