The Author(s) *BMC Bioinformatics* 2016, **17**(Suppl 14):417
DOI 10.1186/s12859-016-1266-4

BMC Bioinformatics

RESEARCH Open Access

An analytical upper bound on the number of loci required for all splits of a species tree to appear in a set of gene trees

Lawrence H. Uricchio[1], Tandy Warnow[2] and Noah A. Rosenberg[1*]

From 14th Annual Research in Computational Molecular Biology (RECOMB) Comparative Genomics Satellite Workshop
Montreal, Canada. 11-14 October 2016

Abstract

Background: Many methods for species tree inference require data from a sufficiently large sample of genomic loci in order to produce accurate estimates. However, few studies have attempted to use analytical theory to quantify "sufficiently large".

Results: Using the multispecies coalescent model, we report a general analytical upper bound on the number of gene trees *n* required such that with probability *q*, each bipartition of a species tree is represented at least once in a set of *n* random gene trees. This bound employs a formula that is straightforward to compute, depends only on the minimum internal branch length of the species tree and the number of taxa, and applies irrespective of the species tree topology. Using simulations, we investigate numerical properties of the bound as well as its accuracy under the multispecies coalescent.

Conclusions: Our results are helpful for conservatively bounding the number of gene trees required by the ASTRAL inference method, and the approach has potential to be extended to bound other properties of gene tree sets under the model.

Keywords: Bipartitions, Coalescent, Gene trees, Species trees

Background

The genomic era presents new challenges for phylogenetic inference studies, because evolutionary processes such as incomplete lineage sorting can cause gene trees at different genomic loci to have different topologies. Many modern techniques therefore capitalize on the independent information available at multiple loci in order to inform phylogenetic estimates (e.g. [1–6]).

One family of phylogenetic methods employs "consensus estimation," in which a set of gene trees on a shared taxon set is used to infer a single consensus species tree that summarizes the information in the input gene tree collection [7, 8]. In a consensus method—and in more general "summary" methods that do not necessarily require the taxon set to be identical across loci—for each of a series of genomic loci, a rooted or unrooted gene tree is first computed, and particular features of the gene tree set are used to compute an estimate of the species tree (e.g. [8–16]).

Consensus and summary methods are often chosen in species tree inference studies because they typically have desirable properties, including computational efficiency, scalability to trees with many taxa, and conceptual simplicity. This latter feature makes such methods suitable for mathematical analyses, many of which have emphasized the property of statistical consistency under a standard model for gene tree evolution, the multispecies coalescent [3]. In a consistent method, as the number of sampled gene trees increases, the probability that the species tree estimate from a random sample of gene trees produced

*Correspondence: noahr@stanford.edu
[1] Department of Biology, Stanford University, Stanford, CA, USA
Full list of author information is available at the end of the article

under the model accords with the true species tree topology approaches 1, irrespective of the species tree topology and branch lengths. Many consensus and summary methods have been shown to be consistent under the multispecies coalescent model (e.g. [8, 10, 11, 13–16]), further justifying their applicability in species tree inference problems.

Mirarab et al. [17] developed one such method: ASTRAL. Given a tree, a *bipartition*, or *split*, corresponds to a cut on one of the branches of the tree, dividing the taxa into two subsets (Fig. 1). Define a gene tree set $\mathcal{G}$ on the same taxon set as the species tree to be a *bipartition cover* of the species tree if for each bipartition in the species tree, at least one gene tree in $\mathcal{G}$ possesses the bipartition. ASTRAL—and the efficiency improvement ASTRAL-II [18]—reports a species tree estimate by searching a space of species trees that draw their bipartitions from a specified input set X. Choosing X to be the set of bipartitions in $\mathcal{G}$ suffices to ensure that ASTRAL is statistically consistent under the multispecies coalescent model [17], because as increasingly many gene trees are included in $\mathcal{G}$, the probability approaches 1 that each bipartition in the true species tree will appear in at least one gene tree, so that $\mathcal{G}$ will be a bipartition cover with probability approaching 1.

How many gene trees are required so that a random set of gene trees is likely to be a bipartition cover of the species tree? For consistent methods, by definition, asymptotically as the number of gene trees increases without bound, the species tree estimate will be accurate with probability 1. However, relatively few analytical recommendations are available for the number of loci required before the probability is high that specified properties of gene tree sets are achieved [8, 19–21]; in the case of ASTRAL, the consistency proof gives no guidance on the number of gene trees required before $\mathcal{G}$ is likely to be a bipartition cover. In place of an analytical treatment, the speed of convergence of consistent methods might typically be examined by simulation-based evaluations (e.g. [10, 22, 23]); although simulations can provide useful insights into the number of required loci, both because they do not produce provable findings and because their parameter choices are inexhaustive, they can have limited generality.

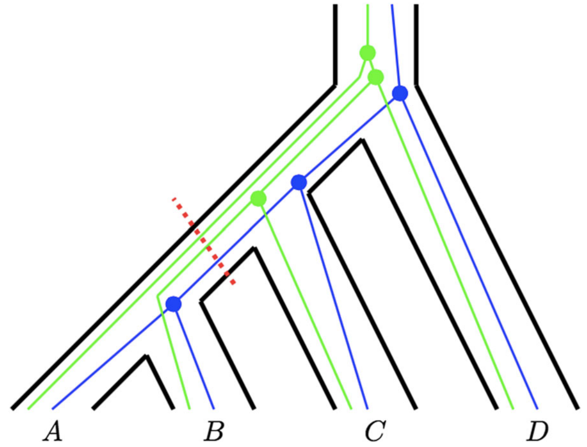

Fig. 1 Schematic of a species tree (*black*) and two gene trees (*blue*, *green*). Coalescent events in a gene tree are constrained to occur only once lineages are present in the same population. The red dashed line indicates a species tree bipartition AB|CD, separating species A and B from species C and D. The same bipartition occurs in the blue gene tree; by contrast, the green gene tree does not contain this bipartition, instead containing AD|BC

Here, we produce a general analytical upper bound for the minimal number of gene trees required for a gene tree set to produce with high probability a bipartition cover of the species tree. As a function of the number of taxa in the species tree, a probability threshold, and a single additional parameter describing the species tree branch lengths, we determine an upper bound on the number of loci needed before the bipartition set represented in a collection of gene trees includes—with the specified minimum probability—all bipartitions in the true species tree. We compare the analytical upper bound to values computed using simulations under the multispecies coalescent model. Our approach can potentially assist in obtaining other, similar upper bounds for the number of loci required before other specific features are likely to appear in gene tree collections.

Results and discussion

Gene tree discordance and the multispecies coalescent

We begin by briefly reviewing the multispecies coalescent model. Under the model, the genealogical history of orthologous lineages from k species is modeled backward in time conditional on a fixed rooted species tree with topology and branch lengths specified. Looking back in time, lineages from a pair of species cannot share common ancestry more recently than the time at which the species share common ancestry (Fig. 1). As a result, conditional on the species tree, not all topologies are equally likely for the gene tree; moreover, a random sample of gene trees that have evolved on the species tree contains information about the species tree topology and branch lengths [24]. In a general treatment of the model, the number of lineages per species is arbitrary, but here we restrict attention to one lineage per species.

Studies of the properties of inference methods applied to sets of gene trees produced under the model can make use of analytical formulas for the probability distribution of gene tree topologies conditional on a species tree [22, 25]. Such formulas employ the species tree topology and branch lengths as parameters, producing a discrete distribution that contains a probability for each possible gene tree topology. This distribution is complex, potentially with significant weight on gene tree topologies that disagree with the species tree, and its properties can differ substantially for species trees with different topologies and different numbers of species [25–28]. In general, under the model, the extent of the disagreement of gene tree topologies with species tree topologies increases as branch lengths in species trees decrease [9, 25], particularly when multiple short branches occur in succession [29].

A key quantity in evaluating gene tree probabilities is a function $g_{i,j}(T)$ that computes the probability that exactly $i-j$ coalescent events happen in time T, beginning from i lineages at time 0 [30]:

$$g_{i,j}(T) = \sum_{k=j}^{i} \frac{e^{-k(k-1)T/2}(2k-1)(-1)^{k-j} j_{(k-1)}\, i_{[k]}}{j!\,(k-j)!\, i_{(k)}}, \tag{1}$$

where $a_{(k)} = a(a+1)\dots(a+k-1)$, $a_{[k]} = a(a-1)\dots(a-k+1)$, and $a_{(0)} = a_{[0]} = 1$. T is measured in coalescent time units, representing a number of generations normalized by the number of gene copies of a locus present in a population ($2N$ for diploids, where N is the effective population size measured as a number of individuals).

Bipartitions

A tree with k leaf nodes has $2k-3$ bipartitions: $k-3$ nontrivial bipartitions in which each of the subsets has at least two leaves, and k trivial bipartitions produced from cuts that separate one leaf from the other $k-1$ leaves. The k trivial bipartitions appear in every tree topology with a fixed leaf label set; henceforth we assume that bipartitions are nontrivial unless otherwise noted. The number of leaves in the larger of the two leaf subsets of a (nontrivial) bipartition is at most $k-2$. The bipartition separating, for example, taxa A and B from taxa C and D, is annotated AB|CD (Fig. 1).

Consider a species tree and a gene tree—both on the same taxon set—in which one gene tree lineage is sampled per species. We say that a nontrivial bipartition ϕ of the species tree is *observed* in the gene tree if for some internal node of the gene tree, a cut on that branch produces the bipartition ϕ of the leaf nodes. For a set $\mathcal{G}$ of gene trees, if each of the $k-3$ nontrivial bipartitions of a species tree S is observed for at least one gene tree in the set, we say that $\mathcal{G}$ is a *bipartition cover* of S.

For gene trees and species trees sharing the same set of k taxa, our goal is to study the probability that a random gene tree set $\mathcal{G}$ containing n gene trees sampled under the multispecies coalescent model is a bipartition cover of a species tree S. We then use this calculation to set an upper bound on the number of loci n required so that with a specified minimum probability, a random n-locus gene tree set is a bipartition cover of S.

Exact computation for four-taxon species trees

We first calculate for four-taxon species trees the exact probability that a gene tree set is a bipartition cover of a species tree. A four-taxon species tree S has only one nontrivial bipartition (Fig. 1), which appears in five of the 15 rooted gene tree topologies. Consider a species tree whose nontrivial bipartition is AB|CD. This bipartition appears in the gene trees with topologies ((AB),(CD)), (((AB),C),D), (((AB),D),C), (((CD),A),B), and (((CD),B,A)).

We compute the probability that a set $\mathcal{G}$ of gene trees is a bipartition cover for a four-taxon species tree S with bipartition AB|CD. Because the species tree has only one nontrivial bipartition, all that is required is for one of the gene trees in $\mathcal{G}$ to have one of the five topologies with the bipartition AB|CD. For four-taxon species trees, it is straightforward to calculate the probabilities under the multispecies coalescent model of each of the 15 gene tree topologies [27]. The probability that a gene tree possesses the species tree bipartition and hence is a bipartition cover is the sum of the probabilities of the five gene tree topologies with bipartition AB|CD.

We must consider two cases, in which S represents the symmetric (Fig. 2a) or asymmetric species tree topology (Fig. 2b). Employing tabulations of gene tree probabilities for four-taxon species trees ([27], Tables 4 and 5), we examine both species tree topologies, denoting the probability that a gene tree has bipartition AB|CD in the symmetric case by P_1^s and in the asymmetric case by P_1^a. The subscript 1 indicates that this quantity is for a single gene tree; we will generalize to sets of n gene trees in the next step. Labeling the species tree branch lengths in coalescent time units by T_1 and T_2 as in Fig. 2, in the symmetric case,

$$\begin{aligned}
P_1^s = &\left[g_{2,1}(T_1)g_{2,1}(T_2) + \frac{1}{3}g_{2,1}(T_1)g_{2,2}(T_2)\right.\\
&\left. + \frac{1}{3}g_{2,2}(T_1)g_{2,1}(T_2) + \frac{1}{9}g_{2,2}(T_1)g_{2,2}(T_2)\right]\\
&+ \left[\frac{1}{3}g_{2,1}(T_1)g_{2,2}(T_2) + \frac{1}{18}g_{2,2}(T_1)g_{2,2}(T_2)\right]\\
&+ \left[\frac{1}{3}g_{2,1}(T_1)g_{2,2}(T_2) + \frac{1}{18}g_{2,2}(T_1)g_{2,2}(T_2)\right]\\
&+ \left[\frac{1}{3}g_{2,2}(T_1)g_{2,1}(T_2) + \frac{1}{18}g_{2,2}(T_1)g_{2,2}(T_2)\right]\\
&+ \left[\frac{1}{3}g_{2,2}(T_1)g_{2,1}(T_2) + \frac{1}{18}g_{2,2}(T_1)g_{2,2}(T_2)\right].
\end{aligned}$$

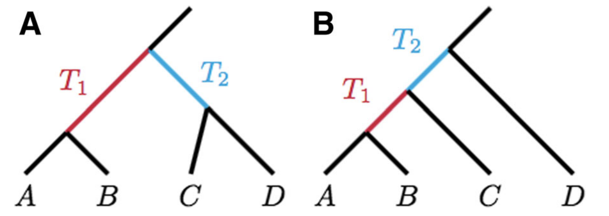

Fig. 2 Two species tree topologies with four taxa. **a** Symmetric topology. **b** Asymmetric topology. Times T_1 and T_2 denote the species tree internal branch lengths

For the asymmetric case,

$$\begin{aligned}P_1^a = &\left[\frac{1}{3}g_{2,1}(T_1)g_{2,2}(T_2) + \frac{1}{9}g_{2,2}(T_1)g_{3,2}(T_2)\right.\\ &\left.+\frac{1}{9}g_{2,2}(T_1)g_{3,3}(T_2)\right]\\ &+\left[g_{2,1}(T_1)g_{2,1}(T_2) + \frac{1}{3}g_{2,1}(T_1)g_{2,2}(T_2)\right.\\ &+\frac{1}{3}g_{2,2}(T_1)g_{3,1}(T_2) + \frac{1}{9}g_{2,2}(T_1)g_{3,2}(T_2)\\ &\left.+\frac{1}{18}g_{2,2}(T_1)g_{3,3}(T_2)\right]\\ &+\left[\frac{1}{3}g_{2,1}(T_1)g_{2,2}(T_2) + \frac{1}{9}g_{2,2}(T_1)g_{3,2}(T_2)\right.\\ &\left.+\frac{1}{18}g_{2,2}(T_1)g_{3,3}(T_2)\right]\\ &+\left[\frac{1}{18}g_{2,2}(T_1)g_{3,3}(T_2)\right] + \left[\frac{1}{18}g_{2,2}(T_1)g_{3,3}(T_2)\right].\end{aligned}$$

The five terms demarcated by brackets in P_1^s and P_1^a give the probabilities of the five gene tree topologies with bipartition AB|CD: ((AB),(CD)), (((AB),C),D), (((AB),D),C), (((CD),A),B), and (((CD),B,A)), respectively.

Simplifying these equations using Eq. 1, we find that

$$P_1^s = 1 - \frac{2}{3}e^{-(T_1+T_2)} \qquad (2)$$

$$P_1^a = 1 - \frac{2}{3}e^{-T_1}. \qquad (3)$$

Note that these two equations are similar in that in each case, the quantity in the exponent, $T_1 + T_2$ or T_1, corresponds to the length of the only internal branch of the unrooted species tree (Fig. 2).

Equations 2 and 3 give the probabilities that a single gene tree is a bipartition cover of the species tree, in the symmetric and asymmetric cases, respectively. Recall that our goal is to calculate the probability that a set $\mathcal{G}$ of n gene trees is a bipartition cover, or that the species tree bipartition is observed in at least one of n sampled gene trees. This quantity—P_n^s in the symmetric case and P_n^a in the asymmetric case—is 1 minus the probability that the bipartition is observed in none of the n trees. Because each gene tree is independent conditional on the species tree, we have

$$P_n^s = 1 - (1 - P_1^s)^n \qquad (4)$$

$$P_n^a = 1 - (1 - P_1^a)^n. \qquad (5)$$

In Fig. 3, we plot P_n^a as a function of the number of loci n for several fixed values of T_1; the behavior of P_n^s is analogous, except with T_1 replaced by $T_1 + T_2$. For each value of T_1, P_n^a increases with n, approaching 1 as $n \rightarrow \infty$. For larger T_1, the initial probability that a single gene tree has bipartition AB|CD is greater, so that the number of gene trees required before P_n^a achieves a specified value is smaller. As $T_1 \rightarrow 0$, gene trees approach a scenario in which the gene lineages from species A, B, and C persist into the common ancestor of the three species. Each possible sequence of coalescences among these three lineages is equally likely, and the probability that a random gene tree contains the nontrivial bipartition AB|CD is $P_1^a = \frac{1}{3}$. P_n then approaches $1 - (\frac{2}{3})^n$.

A general upper bound for *k*-taxon species trees

For $k > 4$, the number of nontrivial bipartitions in a k-taxon species tree exceeds 1, and the event that a random gene tree possesses a nontrivial species tree bipartition ϕ_1 is not independent of the event of its possessing a second such bipartition ϕ_2. To perform a comparably simple calculation in the general k-taxon case to that achieved in the four-taxon case, we focus on deriving a lower bound on the probability that a random n-locus gene tree set $\mathcal{G}$ is a bipartition cover of a k-taxon species tree S.

Let S be a rooted k-taxon species tree with fixed topology and branch lengths. Denote the $k-3$ nontrivial bipartitions of S by $\phi_1, \phi_2, \ldots, \phi_{k-3}$. Denote the $k-2$ internal branches of S by $e_1, e_2, \ldots, e_{k-2}$, with associated lengths $T_1, T_2, \ldots, T_{k-2}$. If one side of the root of S has only a single leaf, then the internal branch immediately

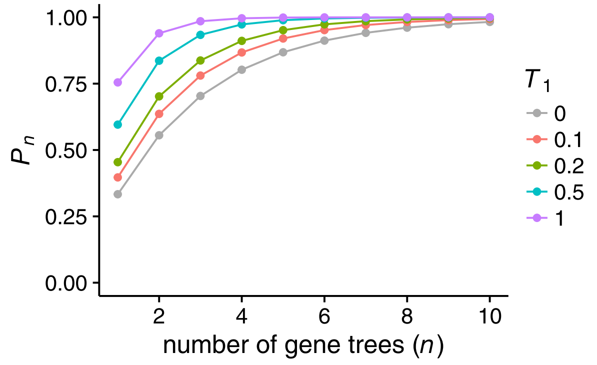

Fig. 3 The probability (P_n) that a random set of n gene trees under the multispecies coalescent is a bipartition cover of a four-taxon asymmetric species tree, as a function of n. Points represent the exact probability computed at each n, for several values of T_1 (Eq. 5)

descended from the other side is associated with a trivial bipartition. We indicate this internal branch by e_c, with $c \in \{1, 2, \ldots, k-2\}$, and we denote its associated branch length T_c. If both sides of the root of S each have at least two descendant leaves, then each of the $k-2$ internal branches is associated with a nontrivial bipartition, and the two branches immediately descended from the root share the same bipartition. We indicate by $\{1, 2, \ldots, k-2\} \setminus c$ the set of indices for internal branches that produce nontrivial bipartitions, understanding that if the two sides of the root each have at least two descendant leaves so that e_c does not exist, this index set reduces to $\{1, 2, \ldots, k-2\}$.

Let $E_{i,n}$ be the event that bipartition ϕ_i is observed at least once in a set $\mathcal{G}$ of n random gene trees, and let $Q_{i,n} = \mathbb{P}[E_{i,n}]$ be the associated probability that at least one of n random gene trees possesses ϕ_i. Then $E_n = E_{1,n} \cap E_{2,n} \cap \cdots \cap E_{k-3,n}$ is the event that a gene tree set $\mathcal{G}$ with n gene trees is a bipartition cover of S. Denote by $Q_n = \mathbb{P}[E_n]$ the probability that a random gene tree set is a bipartition cover: that among n gene trees, all bipartitions of S appear at least once.

The $Q_{i,n}$ have a complex dependence, so that if a gene tree possesses one of the bipartitions ϕ_i, its conditional probability of possessing another bipartition ϕ_j might substantially increase in relation to the unconditional probability. Our strategy for bounding the desired probability Q_n from below amounts to supposing that each bipartition ϕ_i is as improbable as the least-probable bipartition and bounding the probability of the least-probable bipartition from below (Lemma 1). We then disregard the dependence among the $Q_{i,n}$ to bound from below the joint probability that all of the $E_{i,n}$ are observed in a gene tree set (Theorem 2).

Let $T_{\min} = \min_{i \in \{1,2,\ldots,k-2\}} T_i$ denote the length of the shortest internal branch in the species tree S. We obtain a lower bound on $Q_{i,n}$, which we then use to bound Q_n. Our lower bound for Q_n is a function of only k, $T_{\min}$, and n, and it can be inverted to produce an upper bound on the smallest n that achieves a desired minimal value for Q_n.

Lemma 1 $\min_{i \in \{1,2,\ldots,k-3\}} Q_{i,n} \geq 1-[1-g_{k-2,1}(T_{\min})]^n$.

Proof Consider $Q_{i,n}$ for some i. $Q_{i,n}$ is the probability that bipartition ϕ_i is observed in at least one of n random gene trees that are conditionally independent given the species tree. It therefore equals 1 minus the probability that ϕ_i fails to be observed in all n gene trees: $Q_{i,n} = 1-(1-Q_{i,1})^n$. Because for fixed $n \geq 1$, the function $1-(1-x)^n$ increases monotonically in x on $[0, 1]$,

$$\min_{i \in \{1,2,\ldots,k-3\}} Q_{i,n} = \min_{i \in \{1,2,\ldots,k-3\}} [1-(1-Q_{i,1})^n] = 1-\left(1-\min_{i \in \{1,2,\ldots,k-3\}} Q_{i,1}\right)^n. \quad (6)$$

To produce a lower bound on $\min_{i \in \{1,2,\ldots,k-3\}} Q_{i,n}$, it remains to bound $\min_{i \in \{1,2,\ldots,k-3\}} Q_{i,1}$ from below. A sufficient condition for bipartition ϕ_i to be observed in a gene tree is for all the lineages descended from the internal branch e_{ϕ_i} associated with ϕ_i in the species tree to coalesce to a single lineage on that branch. In case ϕ_i is associated with two internal branches—the two immediately descended from the root on opposite sides—it is sufficient for the lineages on one side to coalesce to a single lineage on the internal branch associated with that side. Supposing that k_i is the number of taxa descended in S from branch e_i and T_i is the branch length for e_i, the probability $Q_{i,1}$ that ϕ_i is observed in a single gene tree is therefore bounded below by $g_{k_i,1}(T_i)$, and:

$$1-\left(1-\min_{i \in \{1,2,\ldots,k-3\}} Q_{i,1}\right)^n \geq 1-\left[1-\min_{i \in \{1,2,\ldots,k-2\} \setminus c} g_{k_i,1}(T_i)\right]^n. \quad (7)$$

In this step, although the species tree has $k-3$ nontrivial bipartitions, it has $k-2$ internal branches, one of which possibly produces a trivial bipartition. If cuts on two of the $k-2$ internal branches, say j_1 with k_{j_1} descendant leaf nodes and j_2 with k_{j_2} descendant leaf nodes, produce the same (nontrivial) bipartition ϕ_i, then $Q_{i,1} \geq g_{k_{j_1},1}(T_{j_1})$ and $Q_{i,1} \geq g_{k_{j_2},1}(T_{j_2})$.

The quantity $g_{k_i,1}(T_i)$—the probability that k_i lineages coalesce to 1 lineage during time T_i—decreases monotonically with increasing k_i, and increases monotonically with increasing T_i. Because a species tree internal branch associated with a nontrivial bipartition has at most $k-2$ descendant leaves, and because the shortest internal branch length is $T_{\min}$,

$$g_{k_i,1}(T_i) \geq g_{k-2,1}(T_i) \geq g_{k-2,1}(T_{\min}). \quad (8)$$

This condition applies to each of the $k-2$ internal branches—including both immediately descended from the root in the case that the root does not have a pendant edge as one of its descendants. We take the minimum over internal branches that produce nontrivial bipartitions to obtain

$$\min_{i \in \{1,2,\ldots,k-2\} \setminus c} g_{k_i,1}(T_i) \geq g_{k-2,1}(T_{\min}). \quad (9)$$

We can connect inequalities 6, 7, and 9 to conclude

$$\min_{i \in \{1,2,\ldots,k-3\}} Q_{i,n} \geq 1-[1-g_{k-2,1}(T_{\min})]^n. \quad (10)$$

We thus have the desired result. □

The approach of this proof amounts to replacing the species tree S with $S_{T_{\min}}$, a tree with the same topology as S but with all internal branch lengths set to $T_{\min}$, the minimum branch length in S. Next, it is noted that each bipartition is at least as probable as the least probable bipartition. The probability of the least probable bipartition is then bounded from below by computing a lower bound on one specific way of observing an arbitrary bipartition: the probability of a bipartition is at least as great as the probability that all of the lineages for leaves that descend from its associated internal edge coalesce on that edge.

Now that we have a lower bound for the probability of an arbitrary bipartition, it remains to simultaneously consider all $k-3$ bipartitions.

Theorem 2 $Q_n \geq 1-(k-3)[1-g_{k-2,1}(T_{\min})]^n$.

Proof As the probability of an intersection, Q_n can be written $Q_n = \mathbb{P}[E_n] = \mathbb{P}[\bigcap_{i=1}^{k-3} E_{i,n}]$. The minimal probability of the intersection of a set of possibly dependent events can be bounded by Bonferroni's inequality [31]. It follows that

$$Q_n \geq 1 - \sum_{i=1}^{k-3} \mathbb{P}[\overline{E}_{i,n}], \tag{11}$$

where $\overline{E}_{i,n}$ is the complement of event $E_{i,n}$.

We then have

$$\begin{aligned} Q_n &\geq 1 - \sum_{i=1}^{k-3} (1-\mathbb{P}[E_{i,n}]) \\ &= \left(\sum_{i=1}^{k-3} Q_{i,n} \right) - (k-4) \\ &\geq (k-3) \left(\min_{i \in \{1,2,\ldots,k-3\}} Q_{i,n} \right) - (k-4). \end{aligned} \tag{12}$$

We invoke Lemma 1 to obtain $\min_{i\in\{1,2,\ldots,k-3\}} Q_{i,n} \geq 1-[1-g_{k-2,1}(T_{\min})]^n$, from which

$$Q_n \geq 1-(k-3)[1-g_{k-2,1}(T_{\min})]^n . \tag{13}$$

This completes the proof. □

Note that given the species tree S, for small values of n, it is possible for $(k-3)[1-g_{k-2,1}(T_{\min})]^n \geq 1$, so that the theorem produces a negative value for the lower bound on Q_n. Because Q_n is a probability, in these cases, we have the trivial result $Q_n \geq 0$. As n increases, however, eventually $(k-3)[1-g_{k-2,1}(T_{\min})]^n < 1$, so that in the theorem, Q_n is bounded from below by a positive quantity.

By solving for n, for a specified probability q, Eq. 13 can be used to calculate an upper bound on the minimal value of n for which $Q_n \geq q$. Setting $Q_n = q$ for $0 < q < 1$,

$$n = \frac{\log[(1-q)/(k-3)]}{\log[1-g_{k-2,1}(T_{\min})]}. \tag{14}$$

Equation 14 gives an upper bound on the number of sampled gene trees required for a random gene tree set to be a bipartition cover with probability at least q. It applies irrespective of the species tree topology and branch lengths.

Influences on the upper bound

For fixed values of q, we numerically computed the number of gene trees n required for achieving $Q_n \geq q$ in Eq. 14. In Fig. 4, we plot $\log_{10}(n)$ as a function of the number of taxa k for a range of minimum branch lengths and $q = 1-10^{-2}$ and $q = 1-10^{-5}$.

When $T_{\min} = 1$ or $T_{\min} = 0.5$, so that the shortest internal branch length in the species tree has a value of 1 or 0.5 coalescent time units, n grows slowly as a function of k and remains less than 10^4 for species trees containing up to 30 species. By contrast, when $T_{\min} = 0.2$ or $T_{\min} = 0.1$, species trees with up to $k = 8$ taxa have $n < 10^4$, but the number of gene trees n grows rapidly and exceeds 10^4 for larger k. The patterns are fairly insensitive to the value of q, as q contributes to Eq. 14 only via the logarithmic term $\log(1-q)$.

Accuracy of the upper bound

We next compared our upper bound on the number of loci required to produce a bipartition cover with probability q (Eq. 14) to values of this number of loci obtained in stochastic simulations under the multispecies coalescent. The simulations allow us to quantify the extent to which our upper bound overestimates the true number of required gene trees.

Simulations were conducted using COAL [25] to compute the exact multinomial distribution of gene tree topologies for "caterpillar" species trees in which all branch lengths were set to $T_{\min}$. The caterpillar case represents a difficult scenario for species tree inference, as the extent of gene tree discordance can be greater with caterpillar species trees than other species tree topologies [28, 29, 32, 33]. For fixed values of n_s, the number of simulated gene trees in gene tree sets, we resampled 10^4 independent gene tree sets from this exact multinomial distribution, identifying for each gene tree set all gene tree clades that appeared in at least one of the random gene trees. This clade identification step was conducted using Biopython [34].

Next, we recorded the empirical proportion of simulations in which the n_s gene trees produced a bipartition cover of the species tree. Treating this empirical probability of a bipartition cover as an estimate of Q_{n_s}, we then computed the number of loci n in Eq. 14 using the estimated $\hat{Q}_{n_s}$ for q, denoting this number of loci n_b. The ratio $\frac{n_b}{n_s}$ represents the factor by which our upper bound on the minimum number of loci required for producing a bipartition cover exceeded the actual number of loci required in simulated gene tree sets. A value of $\frac{n_b}{n_s} = 1$ indicates that our upper bound is accurate; values larger than 1

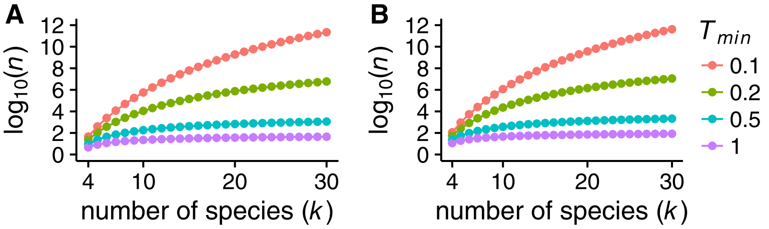

Fig. 4 Upper bound on the number of gene trees required for a random set of n gene trees to have probability at least q of being a bipartition cover of a k-taxon species tree with smallest internal branch length $T_{\min}$. The plot uses Eq. 14. **a** $q = 0.99$. **b** $q = 0.99999$. The maximal number of independent gene trees in a genome is on the order of 10^4 to 10^5

indicate that our upper bound overestimates the number of required gene trees by a factor of $\frac{n_b}{n_s}$.

Figure 5 presents $\frac{n_b}{n_s}$ as a function of q. In each panel, representing different values of $T_{\min}$, $\frac{n_b}{n_s}$ is relatively close to 1 for $k = 4$ taxa, indicating a reasonably accurate upper bound. As k increases, $\frac{n_b}{n_s}$ progressively increases as well. For small k, with relatively few internal branches, fewer ways exist for coalescent events to occur other than on the internal branch of minimum length, so that our consideration of only those coalescences in obtaining the bound disregards fewer alternative ways of producing bipartitions. It hence produces a more accurate n_b.

Comparing the three panels of Fig. 5, we see that $\frac{n_b}{n_s}$ is smaller and the bound n_b is therefore tighter when $T_{\min}$ is large than when $T_{\min}$ is small. For small $T_{\min}$, it is unlikely that all lineages below a species tree branch of length $T_{\min}$ will coalesce on the branch, so that our consideration of only the case in which such coalescences occur in producing Eq. 14 is less accurate. For each $T_{\min}$ value, the level of overestimation does not strongly depend on the value of q, especially for q near 1.

Conclusions

We have derived a general analytical upper bound under the multispecies coalescent on the number of gene trees required for observing with a specified probability q all bipartitions of a species tree. In addition to the number of taxa and the probability q, our upper bound (Eq. 14) depends on a single parameter, the shortest internal branch of the true species tree. This simplicity enables general applicability of a bound that is relatively straightforward to calculate. We find that only a small number of gene trees is required, provided the minimum species tree branch length is not much shorter than the coalescent time scale ($T_{\min} \gtrsim 0.5$). Even when the shortest branch is small relative to the coalescent time scale ($T_{\min} \approx 0.1$), genomic studies of $\approx 10^4$ loci in $k \lesssim 8$ species will produce a bipartition cover of the species tree with high probability. Because our upper bound is a conservative overestimate, it is likely that the bipartition covers useful in the ASTRAL method [17, 18]—which relies on observing all bipartitions of the true species tree in a set of input gene trees—can often be achieved in realistic scenarios with considerably fewer loci.

Species tree branch lengths

Because our upper bound depends on $T_{\min}$, to assess the number of gene trees required for producing bipartition covers in practical studies, we can examine the properties of $T_{\min}$ in models in which not only the gene trees are modeled conditional on fixed species trees, but in which the species trees are modeled as random quantities as well. Stadler & Steel ([35], Theorem 3.3) showed

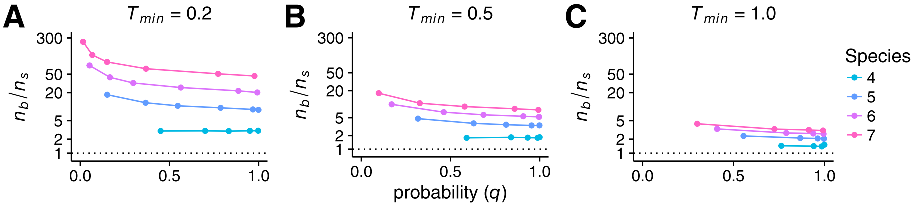

Fig. 5 The ratio $\frac{n_b}{n_s}$ of the upper bound on the minimum number of gene trees required to obtain a bipartition cover with probability q (Eq. 14) to the corresponding number of simulated gene trees required to obtain a bipartition cover with probability q. The ratio is plotted as a function of q, for several values of the number of species k. **a** $T_{\min} = 0.2$. **b** $T_{\min} = 0.5$. **c** $T_{\min} = 1.0$. The y-axis is plotted on a logarithmic scale. Irregular spacing of q values is a result of our simulation procedure, in which each q is determined from 10^4 simulations at a fixed n_s in the set {1, 2, 3, 5, 10, 20, 50, 100, 200, 500}. Note that for some large values of n_s at a fixed $T_{\min}$, all 10^4 simulations produced a bipartition cover, meaning that $\hat{Q}_{n_s} = q = 1$. In these cases, n_b computed from Eq. 14 is infinite and we do not plot $\frac{n_b}{n_s}$

that in the Yule pure birth process for speciation, in which each species lineage speciates forward in time at rate λ, an arbitrary internal branch length has an exponential distribution with rate 2λ. The $k-2$ internal branch lengths in a species tree with k taxa are independent and identically distributed under the model. Hence, $T_{\min}$, as the minimum value of $k-2$ independent exponentially distributed random variables, each with rate 2λ, is exponentially distributed with rate $\sum_{i=1}^{k-2} 2\lambda = 2(k-2)\lambda$. The expected minimum species tree branch length under the Yule model is then $\mathbb{E}[\,T_{\min}] = 1/[\,2(k-2)\lambda]$.

To perform numerical calculations, we chose a range of values of λ on the basis of empirical studies; in the great apes, internal branch lengths of the species tree are consistent with a speciation rate of $\lambda \approx 0.5$ events per coalescent time unit [36, 37], and for primates, Stadler et al. [37] produced an estimate of $\lambda \approx 0.28$. In warblers, Bokma [38] estimated the rate of speciation to be 0.36 per million years. Assuming an effective population size of $N_e = 5 \times 10^4$ and a generation time of 1 year [39], we arrive at $\lambda \approx 0.14$ events per unit of time.

In Fig. 6a, we plot $\mathbb{E}[\,T_{\min}]$ under the Yule model of speciation, as a function of the number of taxa k and the speciation rate λ.When speciation happens rarely relative to the coalescent timescale ($\lambda \leq 0.2$), for up to $k = 15$ species, $\mathbb{E}[\,T_{\min}] \geq 1/(2 \times 13 \times 0.2) \approx 0.19$. When speciation events happen more frequently ($\lambda = 0.5$), however, $\mathbb{E}[\,T_{\min}]$ goes below 0.19 at $k = 8$ species, and $\mathbb{E}[\,T_{\min}] < 0.19$ for $k = 5$ when $\lambda = 1$.

Figure 6b plots the value of n in Eq. 14 that is required to obtain a bipartition cover with probability $q = 0.99$, as a function of the expected minimum branch lengths from Fig. 6a. When speciation is slow ($\lambda \leq 0.2$, e.g. warblers), species trees with $k = 15$ taxa achieve the high probability of 0.99 of producing bipartition covers with a number of gene trees comparable to the scale of the number of independent loci that might be present in a genome ($n = 10^4$ to 10^5). With more frequent speciations, however ($\lambda \geq 0.5$), our upper bound on the required number of gene trees suggests an impractical number of gene trees. Recall that this scenario of large k and small $T_{\min}$ is precisely the case in which our upper bound is most conservative (Fig. 5), so that a stricter upper bound might indicate that the true required number of gene trees is in fact in a range that is practicable in principle.

Extensions

Our analysis of the effect of the speciation rate λ on the number of gene trees required for observing a bipartition cover highlights both the utility and the limitations of our approach. The results apply irrespective of the number of species and the species tree topology and branch lengths; however, to obtain this generality, we have relied on approximations that make our bound conservative. To compute the probability that a gene tree set is a bipartition cover, in Lemma 1, we have assumed that each bipartition is only as probable as the least likely bipartition. Further, considering only the least likely bipartition has amounted to assuming that all branches have the same length as the shortest branch. We have also used a conservative lower bound for the probability of the least likely bipartition. In Theorem 2, we have conservatively assumed that the presence in a gene tree of one species tree bipartition does not affect the presence of another bipartition. By incorporating more parameters for the species tree rather than only the number of species and $T_{\min}$, each of these assumptions can potentially be relaxed to produce a more accurate upper bound on the number of gene trees required for obtaining a bipartition cover.

For example, consider our lower bound for the probability of the least likely bipartition, which assumes that $k-2$ lineages coalesce to a single lineage on the shortest species tree internal branch. Most species trees have no internal branch from which $k-2$ species descend; further, it is unlikely that if such a branch does exist that it is the shortest internal branch. Even in this scenario, many ways exist for the bipartition to be realized by a gene tree other than by all $k-2$ lineages coalescing on the shortest branch.

With the species tree branch lengths and topology taken into account, we can in fact calculate the probability of the least likely bipartition. Suppose a bipartition ϕ of the species tree separates the k taxa into two species groups,

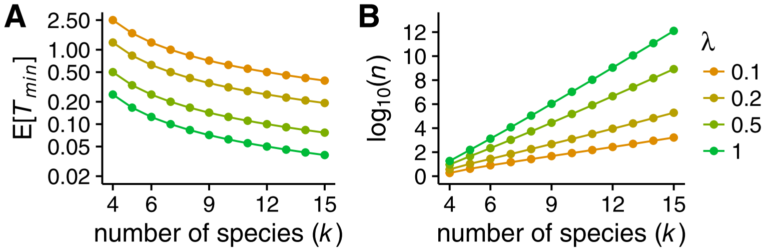

Fig. 6 $T_{\min}$ under the Yule pure birth process for speciation at rate λ speciation events per coalescent time unit. **a** $\mathbb{E}[\,T_{\min}]$ as a function of the number of species k. The y-axis is plotted on a logarithmic scale. **b** The number of gene trees n required in Eq. 14 for obtaining with probability q all species tree bipartitions in a gene tree set, as a function of $\mathbb{E}[\,T_{\min}]$ values from **a**. The value of q is fixed at 0.99. Note that the maximal number of independent gene trees in a genome is approximately 10^4 to 10^5

$\mathcal{T}_\phi$ and $\overline{\mathcal{T}}_\phi$. The probability that bipartition ϕ is observed in a gene tree is then the same as the probability that the gene lineages of the species in either $\mathcal{T}_\phi$ or $\overline{\mathcal{T}}_\phi$ (or both) are monophyletic:

$$\mathbb{P}[E_{\phi,1}] = P_M(\mathcal{T}_\phi) + P_M(\overline{\mathcal{T}}_\phi) - P_{RM}(\mathcal{T}_\phi, \overline{\mathcal{T}}_\phi), \qquad (15)$$

where P_M is the probability of monophyly of a set of gene lineages, P_{RM} is the probability of reciprocal monophyly of a pair of sets of gene lineages, and $\mathbb{P}[E_{\phi,1}]$ is the probability that the bipartition ϕ is observed in a random gene tree (by abuse of notation, we identify the gene lineages of species set $\mathcal{T}_\phi$ with $\mathcal{T}_\phi$, and similarly for $\overline{\mathcal{T}}_\phi$). Recently, Mehta et al. [40] derived formulas for P_M and P_{RM} for arbitrary gene lineage sets conditional on arbitrary fixed species trees with topology and branch lengths specified; using these formulas, it would be possible to exactly calculate the probabilities of each of the $k-3$ bipartitions, and to replace our lower bound on the probability of the least likely bipartition in Lemma 1 with the *exact* minimum.

We note than in addition to ASTRAL, other methods (including in problems with gene duplication and loss rather than incomplete lineage sorting [41]) employ similar constrained search algorithms relying on bipartitions. Some methods have the property that if the input gene tree set is a bipartition cover of the species tree, the true species tree lies in the search space and is feasible to produce as an estimate [12, 42]. Our work thus provides guidance on the maximum number of loci required before the true species tree enters the search space. As a calculation applicable to arbitrary species trees, considering single features and then examining their joint probability by use of a Bonferroni inequality, our approach might thus be applicable in other problems that require a lower bound on the probability that a property is achieved by a gene tree set, or an upper bound on the number of gene trees required for achieving the property. Though it disregards detailed information that might be available about the species tree, the generality of the approach has potential to provide useful bounds on probabilities that are otherwise difficult to evaluate.

Methods

The methods are described throughout the Results and discussion section.

Acknowledgments
We acknowledge support from NSF grants DBI-1146722, DBI-1458059, and DBI-1461364, from NIH grant R01 GM117590, and from a Stanford Center for Computational, Evolutionary, and Human Genomics postdoctoral fellowship to LHU. We thank David Bryant for discussions and Rohan Mehta for computational assistance.

Declarations
Publication charges for this article were funded by NIH grant R01 GM117590 and the corresponding author's departmental sources.
This article has been published as part of *BMC Bioinformatics* Vol 17 Suppl 14, 2016: Proceedings of the 14th Annual Research in Computational Molecular Biology (RECOMB) Comparative Genomics Satellite Workshop: bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-17-supplement-14.

Availability of data and materials
Not applicable.

Authors' contributions
LHU, TW, and NAR designed the project. LHU performed the computational experiments. LHU, TW, and NAR wrote the paper. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1] Department of Biology, Stanford University, Stanford, CA, USA. [2] Departments of Computer Science & Bioengineering, University of Illinois, Urbana-Champaign, IL, USA.

Published: 11 November 2016

References

1. Ané C, Larget B, Baum DA, Smith SD, Rokas A. Bayesian estimation of concordance among gene trees. Mol Biol Evol. 2007;24(2):412–26.
2. Rannala B, Yang Z. Phylogenetic inference using whole genomes. Annu Rev Genomics Hum Genet. 2008;9:217–31.
3. Degnan JH, Rosenberg NA. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol. 2009;24(6): 332–40.
4. Heled J, Drummond AJ. Bayesian inference of species trees from multilocus data. Mol Biol Evol. 2010;27(3):570–80.
5. Knowles LL, Kubatko LS, (eds). Estimating Species Trees: Practical and Theoretical Aspects. New York: Wiley; 2010.
6. Liu L, Xi Z, Wu S, Davis CC, Edwards SV. Estimating phylogenetic trees from genome-scale data. Ann N Y Acad Sci. 2015;1360(1):36–53.
7. Bryant D. A classification of consensus methods for phylogenetics. In: Janowitz MF, Lapointe F-J, McMorris FR, Mirkin B, Roberts FS, editors. Bioconsensus. Providence: American Mathematical Society; 2003. p. 163–83.
8. Degnan JH, DeGiorgio M, Bryant D, Rosenberg NA. Properties of consensus methods for inferring species trees from gene trees. Syst Biol. 2009;58(1):35–54.
9. Degnan JH, Rosenberg NA. Discordance of species trees with their most likely gene trees. PLoS Genet. 2006;2(5):e68.
10. Liu L, Yu L, Pearl DK, Edwards SV. Estimating species phylogenies using coalescence times among sequences. Syst Biol. 2009;58(5):468–77.
11. Liu L, Yu L, Pearl DK. Maximum tree: a consistent estimator of the species tree. J Math Biol. 2010;60(1):95–106.
12. Than C, Nakhleh L. Species tree inference by minimizing deep coalescences. PLoS Comput Biol. 2009;5(9):e1000501.
13. Mossel E, Roch S. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. IEEE/ACM Trans Comput Biol Bioinforma. 2010;7(1):166–71.
14. Liu L, Yu L. Estimating species trees from unrooted gene trees. Syst Biol. 2011;60(5):661–7.
15. Jewett EM, Rosenberg NA. iGLASS: an improvement to the GLASS method for estimating species trees from gene trees. J Comput Biol. 2012;19(3):293–315.
16. Allman ES, Degnan JH, Rhodes JA. Species tree inference by the STAR method and its generalizations. J Comput Biol. 2013;20(1):50–61.
17. Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics. 2014;30(17):541–8.

18. Mirarab S, Warnow T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics. 2015;31(12):44–52.
19. Wu C-I. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics. 1991;127(2):429–35.
20. Roch S. An analytical comparison of multilocus methods under the multispecies coalescent: the three-taxon case. Pac Symp Biocomput. 2013;18(1):297–306.
21. Dasarathy G, Nowak R, Roch S. Data requirement for phylogenetic inference from multiple loci: a new distance method. IEEE/ACM Trans Comput Biol Bioinforma. 2015;12(2):422–32.
22. Wu Y. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution. 2012;66(3):763–75.
23. DeGiorgio M, Degnan JH. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Syst Biol. 2014;63(1):66–82.
24. Allman ES, Degnan JH, Rhodes JA. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J Math Biol. 2011;62(6):833–62.
25. Degnan JH, Salter LA. Gene tree distributions under the coalescent process. Evolution. 2005;59(1):24–37.
26. Pamilo P, Nei M. Relationships between gene trees and species trees. Mol Biol Evol. 1988;5(5):568–83.
27. Rosenberg NA. The probability of topological concordance of gene trees and species trees. Theor Popul Biol. 2002;61(2):225–47.
28. Rosenberg NA, Tao R. Discordance of species trees with their most likely gene trees: the case of five taxa. Syst Biol. 2008;57(1):131–40.
29. Rosenberg NA. Discordance of species trees with their most likely gene trees: a unifying principle. Mol Biol Evol. 2013;30(12):2709–713.
30. Tavaré S. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor Popul Biol. 1984;26(2): 119–64.
31. Rice JA. Mathematical Statistics and Data Analysis. Belmont: Duxbury Press; 1995.
32. Degnan JH. Anomalous unrooted gene trees. Syst Biol. 2013;62(4):574–90.
33. Degnan JH, Rhodes JA. There are no caterpillars in a wicked forest. Theor Popul Biol. 2015;105(1):17–23.
34. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–3.
35. Stadler T, Steel M. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. J Theor Biol. 2012;297(1):33–40.
36. Schrago CG. The effective population sizes of the anthropoid ancestors of the human–chimpanzee lineage provide insights on the historical biogeography of the great apes. Mol Biol Evol. 2014;31(1):37–47.
37. Stadler T, Degnan JH, Rosenberg NA. Does gene tree discordance explain the mismatch between macroevolutionary models and empirical patterns of tree shape and branching times? Syst Biol. 2016;65(4):628–39.
38. Bokma F. Bayesian estimation of speciation and extinction probabilities from (in) complete phylogenies. Evolution. 2008;62(9):2441–5.
39. Milot E, Gibbs HL, Hobson KA. Phylogeography and genetic structure of northern populations of the yellow warbler (*Dendroica petechia*). Mol Ecol. 2000;9(6):667–81.
40. Mehta R, Bryant D, Rosenberg NA. The probability of monophyly of a sample of gene lineages on a species tree. Proc Natl Acad Sci U S A. 2016;113(29):8002–9.
41. Hallett MT, Lagergren J. New algorithms for the duplication-loss model In: Shamir R, Miyano S, Istrail S, Pevzner P, Waterman M, editors. RECOMB 2000: Proceedings of the Fourth Annual International Conference on Computational Biology. New York: ACM Press; 2000. p. 138–46.
42. Yu Y, Warnow T, Nakhleh L. Algorithms for MDC-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. J Comput Biol. 2011;18(11):1543–59.