Published in final edited form as:

Mol Phylogenet Evol. 2017 November ; 116: 13–24. doi:10.1016/j.ympev.2017.07.017.

Inferring rooted species trees from unrooted gene trees using approximate Bayesian computation

Ayed R. A. Alanzi^{1,2} and James H. Degnan¹

¹Department of Mathematics and Statistics, University of New Mexico, 311 Terrace NE, Albuquerque, NM 87131-001, New Mexico, USA

²Department of Mathematics, College of Science and Human Studies at Hotat Sudair, Majmaah University, P.O Box 66, Majmaah 11952, KSA

Abstract

Methods for inferring species trees from gene trees motivated by incomplete lineage sorting typically use either rooted gene trees to infer a rooted species tree, or use unrooted gene trees to infer an unrooted species tree, which is then typically rooted using one or more outgroups. Theoretically, however, it has been known since 2011 that it is possible to consistently infer the root of the species tree directly from unrooted gene trees without assuming an outgroup. Here, we use approximate Bayesian computation to infer the root of the species tree from unrooted gene trees assuming the multispecies coalescent model. It is hoped that this approach will be useful in cases where an appropriate outgroup is difficult to find and gene trees do not follow a molecular clock. We use approximate Bayesian computation to infer the root of the species tree from unrooted gene trees. This approach could also be useful when there is prior information that makes a small number of root locations plausible in an unrooted species tree.

Graphical abstract

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Supplementary materials

Example scripts, and necessary code in C and R used to implement the algorithms in the paper are provided during the review phase online at www.math.unm.edu/~james/SupplementABC.zip. (This will be moved to an appropriate place online if published.)

Keywords

multispecies coalescent; outgroup; midpoint rooting; molecular clock; identifiability; sufficiency

1. Introduction

Methods that infer species trees from estimated gene trees have been developing rapidly over the last 10 years or so. Such methods have included majority and extended majority rule consensus (Gadagkar *et al.*, 2005; Degnan *et al.*, 2009), the minimize deep coalescence and shallowest coalescent methods (Maddison and Knowles, 2006), rooted triple consensus (Ewing *et al.*, 2008), BUCKy (Ané *et al.*, 2007; Larget *et al.*, 2010), STEAC (Liu *et al.*, 2009), STAR (Liu *et al.*, 2009), GLASS/Maximum Tree (Kubatko *et al.*, 2009; Liu *et al.*, 2010b; Mossel and Roch, 2010), MP-EST (Liu *et al.*, 2010a), Matrix Representation with Parsimony (MRP) and other supertree methods (Kupczok *et al.*, 2010), triplet-MRP (Wang and Degnan, 2011), NJ_{st} (Liu and Yu, 2011), approximate Bayesian computation (Fan and Kubatko, 2011) and ASTRAL (Mirarab *et al.*, 2014). Considering this timeline, there has been a clear shift from methods that use rooted gene trees as input (most methods from 2005 to 2010–2011) to those using unrooted gene trees as input (most of the methods listed after 2010). In some cases, later methods can be viewed as unrooted versions of earlier rooted methods, such as an unrooted version of the minimize deep coalescence criterion (MDC) (Yu *et al.*, 2011) and pseudolikelihood using quartets instead of rooted triples (Solís-Lemus and Ané, 2016). The method NJ_{st} (Liu and Yu, 2011) can also be viewed as an unrooted version of the earlier STAR method (Liu *et al.*, 2009; Allman *et al.*, 2016). We note that the unrooted version of MDC implemented in PhyloNet (Than *et al.*, 2008) can return a rooted species tree from unrooted gene trees given as input. We discuss this method in comparison to ABC at the end of the paper.

Methods using unrooted trees are convenient in part because fast likelihood programs such as RAxML (Stamatakis, 2006) and PHYML (Guindon *et al.*, 2010) only estimate unrooted trees without a clock. In addition, rooting gene trees could possibly lead to systematic errors above and beyond the usual errors due to gene tree estimation error caused by short alignments or incorrect substitution models. In particular, even if a taxon is an outgroup at the species tree level, it is not necessarily always an outgroup for each gene tree. For

example, the simulations of Huang and Knowles (2009) found that with 6 coalescent units (i.e., $6N$ generations) separating the root of the ingroup taxa and the root of the species tree including the outgroup, there was only a 95% probability of monophyly for the ingroup taxa, meaning that 5% of gene trees rooted using an outgroup would be incorrectly rooted.

A theoretical result from Allman *et al.* (2011b) is that when there are five or more taxa, the true distribution of unrooted gene tree topologies can be used to infer the *rooted* species tree. This result is based on knowing the probabilities of each unrooted gene tree topology, not on any information in the sequence data. In practice, unrooted gene trees are estimated from sequence data from a finite number of loci, and therefore topology probabilities are only estimated. It is therefore not clear how easy it is to estimate the rooted species tree from estimated unrooted gene trees and from finite numbers of loci.

Part of the argument for the identifiability of the rooted species tree from unrooted gene trees is that there are certain inequalities that hold in the gene tree probabilities. For example, consider distinguishing the two rooted species trees $((((a, b), c), d), e)$ versus $((((a, b), c), e), d)$. Both species trees have the same unrooted topologies but imply different inequalities in some of the unrooted gene tree topology probabilities. For the first species tree, lineage c is more likely to coalesce with d than e ; consequently, for this species tree the unrooted gene tree $((a, b), e, (c, d))$ is more probable than gene tree $((a, b), d, (c, e))$. However, for species tree $((((a, b), c), e), d)$, these inequalities are reversed. Consequently, observing more gene trees with topology $((a, b), e, (c, d))$ than topology $((a, b), d, (c, e))$ gives evidence favoring the first species tree over the second species tree. Interestingly, the frequency of the matching unrooted gene tree $((a, b), c, (d, e))$ is not helpful for distinguishing the two species trees. Instead, it is information in the nonmatching unrooted gene trees that is useful for inferring the rooted species tree.

The result in Allman *et al.* (2011b) doesn't lead directly to a method for inferring the rooted species tree. In this paper, we use a modified version of the ABC-inspired algorithm of Fan and Kubatko (2011) to infer the root of the unrooted species tree assuming that the unrooted species tree is known (or has been inferred from some other method).

In Fan and Kubatko (2011), a set of gene trees is used as input data. Species trees are inferred by simulating data sets similar to the input data set from species trees drawn from a prior distribution. A distance is then calculated between the simulated species tree and the observed data. Species trees that had the smallest distance are saved and used to estimate the posterior distribution for the species tree. The idea is that randomly chosen species trees that are relatively close to the true species tree should generate simulated data sets that are more similar to the observed data than species trees that are further from the true species tree.

This method is similar to approximate Bayesian computation (ABC) except that in ABC methods, a data set is simulated from the randomly chosen parameters from the prior, and the distance is computed between observed and simulated data sets. In Fan and Kubatko (2011), instead of simulating gene trees from the species tree chosen from the prior, the expected numbers of gene tree topologies were computed analytically using the distribution of gene trees determined from the multispecies coalescent using the software COAL.

(Degnan and Salter, 2005). This approach was criticized by Buzbas (2012) for not being a genuine ABC method because data sets were not simulated.

We modify the approach of Fan and Kubatko (2011) by using the usual ABC approach of simulating a data set for each parameter (species tree) chosen from the prior. In addition, we use a prior that only consists of rooted species trees with the same unrooted topology, and we use unrooted rather than rooted gene tree topologies. Although rooted species trees are chosen from the prior, the simulated gene trees from each species tree are treated as unrooted when calculating the distance between observed and simulated sets of gene trees.

Distances between datasets are computed using a summary statistic from both the observed and simulated data sets. Species trees simulated from the prior that corresponded to small distances are then used as a basis for an estimated posterior distribution for the parameter (i.e., the species tree). Generally in Bayesian statistics, the goal is to determine the posterior distribution of the parameter given the data. In ABC, we estimate the posterior distribution given that the summary statistic is close to the data, which it is hoped approximates the posterior distribution given the data (Sisson *et al.*, 2007; Joyce and Marjoram, 2008). For this reason, it is desirable to use sufficient statistics as the summary statistics (summaries of the data that do not lose any information useful for doing inference) (Casella and Berger, 2002).

We hope that the ABC approach will be useful in situations where the molecular clock does not hold and an appropriate outgroup is difficult to find, such as in the study of rooting methods for *Orcuttia* by Boykin *et al.* (2010). The method is likely to be impractical unless there is a small number of candidate species trees. For our simulation study, we imagine a case where the unrooted species tree is inferred using an appropriate method, such as ASTRAL, NJ_{st}, BUCKy, or SVDquartets (Chifman and Kubatko, 2014), and the ABC method is used to narrow down plausible locations for the root of the species tree.

Details of our ABC algorithm are given in the methods section below. We perform simulations using five-, eight-, and 16-taxon trees to illustrate the method for different species tree topologies and branch lengths. We also apply the method to an empirical data set where the rooting of *Amborella* with respect to angiosperms has been controversial.

2. Methods

2.1. ABC

The general methodology for ABC methods is to first simulate from the prior distribution for the parameter (in this case, a species tree with branch lengths), then to simulate data from the parameter (gene trees from species trees), and to record a distance between the simulated data set and the observed data. In this project, the observed data as well as simulated data consist of unrooted gene tree topologies. The distance computed between the observed data and the simulated data depends on a choice of summary statistic, and a few variations were used. We apply the method to trees with five, eight, and 16 taxa.

For five taxa, there are only 15 unrooted gene tree topologies, and it is possible to record how often each of the 15 gene tree topologies occurs for both the observed and simulated data. A vector of topology counts therefore is a summary of the data that preserves all relevant information in the sample. A variation is to instead record the splits that occur in a sample of trees, counting how often each split occurs, but discarding information about which split came from which tree. For five taxa, there are 10 possible splits.

For eight taxa, there are 10,395 possible unrooted tree topologies. Having a vector of length 10,395 to record counts of how often each possible topology occurs is impractical, and is an example of the difficulty of using ABC in high-dimensional problems. Fan and Kubatko (2011) deal with this problem for 8 taxa by only recording counts of the rooted gene trees occurring in the observed data, and then counting the number of gene tree topologies in the simulated data that correspond to one of the input trees. However, the number of distinct tree topologies can still be quite high, in some cases with each gene tree having a unique topology (Salichos and Rokas, 2013), and Fan and Kubatko (2011) found their approach to be less accurate for trees with eight tips than for four, which we speculate might be due to the dimensionality problem. For 8 taxa, we instead use only splits, and instead of recording counts of all possible splits, we record the symmetric difference between the set of splits in the observed data versus the simulated data. Details of these algorithms are given below.

The data is composed of counts of G unrooted topologies presented in the observed data. The present work records $n.topo.obs$, which is a vector of the number of times each possible unrooted gene tree topology was observed. For 5 taxa, there are 15 possible unrooted topologies. Thus, $n.topo.obs$ is a vector of length 15, and $n.topo.obs_i$ counts the number of times the i th topology occurs in the data, with $\sum_{i=1}^{15} n.topo.obs_i = G$. Similarly, for the j th simulated data set, there is a vector of counts, and we use $n.topo.sim_{ij}$ to indicate the i th entry of the j th vector of observed counts. For each j , $\sum_{i=1}^{15} n.topo.sim_{ij} = G$. The ABC algorithm for inferring the rooted topology of the species tree involves steps (2)–(7). Step 1 is used to simulate the observed data. We use $T_n = (2n - 5)!!$ to denote the number of unrooted topologies for n taxa.

Algorithm 1

-
- (1) Simulate the observed data by using the program Hybrid-Lambda to simulate from the species tree, which is called $n.topo.obs$.
 - (2) Start with $j = 1$.
 - (3) Simulate a species tree from the prior distribution of rooted species trees.
 - (4) Simulate gene trees from the species tree sampled from the prior by using the program Hybrid-Lambda; the vector of gene tree counts is called $n.topo.sim_j$.
 - (5) Calculate $D_{j.topo} = \sum_{i=1}^{T_n} (n.topo.obs_i - n.topo.sim_{ij})^2$
 - (6) Increment j by 1 and repeat steps (2) – (5) J times.
 - (7) Take the smallest aJ values from step 5, then retain the species trees corresponding to these smallest distances. These trees estimate the posterior distribution.
-

Here α is chosen to be a small number so that only those simulated species trees with simulated data close to the observed data are retained. For this study, we used $\alpha = 0.002$ and $\alpha J = 100$ so that the posterior is formed from the 100 best species trees. Another approach in ABC is to accept samples that have a small distance, less than some chosen tolerance $\delta > 0$. We follow Fan and Kubatko (2011) in accepting a fixed number of trees instead that had the smallest distances. This approach is equivalent to using a tolerance that is based on a quantile of the simulated distribution of distances and is a common approach in practice in ABC (e.g., Beaumont *et al.*, 2002). The fixed tolerance approach leads to a random number of iterations needed to obtain a fixed number of accepted parameters to estimate the posterior distribution. The approach of using the best $100 \times \alpha\%$ of the species trees instead leads to a fixed number J of datasets to simulate and is easier to use in practice because the length of the simulation is known in advance, making it easier to plan the length of time needed for computation.

The present work infers the rooted topology of the species tree by summarizing the posterior distribution of the species tree topologies. For species trees with a large number of taxa, it is possible to use a consensus tree for the topology estimate.

The same method that was implemented for the topology counts can instead use split frequencies of the gene trees. The data is summarized by the counts of splits in the G unrooted gene trees for five taxa presented in the observed data. The vector of split counts is recorded as $n.split.obs$, where $n.split.obs_i$ is the number of times the i th possible split is observed in the data. We use S_n to denote the number of possible nontrivial splits on n taxa. For 5 taxa, there are 10 possible splits. Thus, for five taxa, $n.split.obs$ is a vector of length 10 while $n.topo.obs$ is a vector of length 15. The vector $n.split.obs$ is therefore a lower-dimensional summary of the data than $n.topo.obs$. Although for five taxa, the difference isn't dramatic, the number of possible splits grows much more slowly. For example for $n = 8$ taxa, there are $T_8 = 10,395$ unrooted topologies but only 119 possible splits. In both cases, the distance calculated is the squared Euclidean distance between the count vectors. Each gene tree contributes $n - 3$ nontrivial splits, and therefore $\sum_{i=1}^{S_n} n.split.obs_i = (n - 3)G$, which would also tend to make the counts in $n.split.obs$ larger than the counts in $n.topo.obs$. The ABC algorithm, depending on the split counts, involves steps (2)–(7) below.

Algorithm 2

-
- (1) Simulate the observed data by using the program Hybrid-Lambda to simulate gene trees, and count the splits in the gene trees using $n.split.obs$.
 - (2) Start with $j = 1$.
 - (3) Simulate a species tree from the prior distribution of rooted species trees.
 - (4) Simulate gene trees from the species trees sampled from the prior distribution by using the program Hybrid-Lambda. The count of the splits is called $n.split.sim_j$.
 - (5) Calculate $D_{j,split} = \sum_{i=1}^{S_n} (n.split.obs_i - n.split.sim_{ij})^2$
 - (6) Increment j by 1 and repeat steps (2) – (5) J times.

(7) Take the smallest αJ values from step 5, then retain the species trees corresponding to these smallest distances. Those trees estimate the posterior distribution.

The above two algorithms were applied for trees with five taxa. For trees with eight taxa, splits are recorded as multisets, which are sets that keep track of the multiplicity of each element—the number of times each element of the set occurs. As an example of applying the multiset idea, Figure 1 gives a toy example in which two sets of three trees are shown, and the resulting multisets of splits and their distance is computed. The distance $D = |S_{obs} \setminus S_{sim}| + |S_{sim} \setminus S_{obs}|$ is the sum of the sizes of the set differences.

Algorithm 3

- (1) Extract splits from the observed gene trees into the multiset S_{obs} .
- (2) Start with $j = 1$.
- (3) Simulate a species tree from the prior distribution of rooted species trees.
- (4) Simulate gene trees from the species trees sampled from the prior distribution by using the program Hybrid-Lambda. Extract all splits from the simulated gene trees into the multiset S_{sim} .
- (5) Let $D = |S_{obs} \setminus S_{sim}| + |S_{sim} \setminus S_{obs}|$, the symmetric difference between the two sets of splits.
- (6) Increment j by 1 and repeat steps (2) – (5) J times.
- (7) Take the smallest αJ values from step 5, then retain the sample species trees corresponding to these smallest distances. Those trees estimate the posterior distribution.

In step 5, the distance is the symmetric distance between the two multisets. We note that if the sets S_{obs} and S_{sim} were based on one observed and one simulated tree, respectively, then D would reduce to the Robinson-Foulds (RF) distance (Robinson and Foulds, 1981). The value of D we compute in Algorithm 3 generalizes the RF distance when there are two sets of trees rather than two individual trees. The present work summarizes the posterior distribution of the species tree splits to infer the rooted species tree using majority-rule consensus. To implement all calculations, this study employs a number of scripts, including code in R (Ihaka and Gentleman, 1996). Details are given in the online supplemental files.

This study is similar to Fan and Kubatko (2011) in terms of estimating species trees based on steps (2) to (4) in Algorithms 1–3. However, they compute expected counts of gene trees theoretically, whereas we simulate gene trees in order to use a genuinely ABC algorithm. This is also motivated by the fact that determining the expected counts requires computing an entire gene tree distribution, which is computationally expensive for larger trees (Degnan and Salter, 2005; Wu, 2012). Second, the way that they computed the distance between the observed gene tree and the simulated gene tree differs from ours. Using $n_{obs,i}$ and $n_{exp,i}$ for the observed and expected number of loci, respectively, with the i th topology, they compute the distance as $D_j = \sum_{i=1} (n_{obs,i} - n_{exp,i})^2 / n_{exp,i}$ (summing over possible rooted rather than unrooted topologies), whereas we don't divide by the expected counts (see step (5) in algorithms 1–3). This is largely because when simulated counts are used rather than expected counts, the denominator is often 0. A few additional simulations using alternative definitions of distance found very little difference in results with different distances.

Examples tried included $D_{j,\text{split}} = \sum_{i=1}^{S_n} \frac{(n.\text{split}.obs_i - n.\text{split}.sim_{ij})^2}{1 + n.\text{split}.sim_{ij}}$ and

$D_{j,\text{split}} = \sum_{i=1}^{S_n} |n.\text{split}.obs_i - n.\text{split}.sim_{ij}|$ (Data not shown). Third, they inferred the rooted species tree directly from rooted gene trees, but in this study rooted species trees are estimated by using the unrooted gene trees. Finally, the priors we used were restricted to locating the root of the species tree on an unrooted tree. In their case, the prior for the species tree was based on trees that were short distances from the input gene trees.

2.2. Simulation

2.2.1. Simulation of Five-Taxon Trees—A set of 14 rooted bifurcating trees was selected by using different branch length combinations for three tree shapes: a caterpillar tree, a pseudocaterpillar tree, and a balanced tree (meaning three taxa on one side of the root, and two taxa on the other) to simulate the data. The tree shapes are given in Fig. 2, and the branch length combinations are shown in Table 1. We refer to these species trees as $S_{c,1}, \dots, S_{c,5}$ for the caterpillar shape, $S_{p,1}, \dots, S_{p,4}$ for the pseudocaterpillar shape, and $S_{b,1}, \dots, S_{b,5}$ for the balanced shape. The species tree is the parameter, and the data consist of the unrooted gene tree topologies. Specific branch length combinations used were based on either having all short branches (0.1 coalescent units), all relatively long branches (1.0 coalescent units), or two short and one long branch. The first step involves simulating the data and recording them as counts of either distinct topologies or splits. The vector of observed topology counts simulation is called $n.\text{topo}.obs$. Similarly, the splits of the gene trees are calculated and are called $n.\text{split}.obs$.

The second step is simulating species trees (the parameters) by using a prior. The simulated species trees are then used to simulate data in the form of unrooted gene trees. A uniform prior was used for the species tree topologies. The prior consists of the seven possible rooted topologies of the unrooted tree topology $((a, b), c), d, e$, and are labeled $T_1, \dots, T_7$ (Fig. 2). Here $T_1, \dots, T_4$ have a caterpillar topology with T_1 matching the topology of the model species trees $S_{c,1}, \dots, S_{c,5}$. Tree T_5 has a pseudocaterpillar topology and matches $S_{p,1}, \dots, S_{p,4}$, and T_6 has a balanced topology and matches species trees $S_{b,1}, \dots, S_{b,5}$. An exponential distribution with rate 1.0 coalescent units was used for branch lengths of the species trees, including external branch lengths. A branch length of 1.0 coalescent units predicts a probability of $1 - e^{-1} \approx 0.63$ probability that two lineages coalesce within that branch, thus allowing both coalescence and lack of coalescence to have moderate probabilities. The exponential prior is used so that a wide range of coalescent are simulated. Once the data were simulated with those priors, the topology counts and split counts were recorded.

Gene trees were simulated from species trees using Hybrid-Lambda (Zhu *et al.*, 2015). The simulation used $J = 50,000$. For each repetition of J , a sample size of 100 genes was used. The value of α was set to 0.002 so that the 100 best species trees were retained. The formula αJ was used to determine the desired number of species trees based on the smallest D_j . The value of J was determined by a pilot simulation study using species tree $S_{c,1}$ which suggested that $J = 50,000$ and 100 loci yielded more accurate results than $J = 10,000$ and 500 loci, although the computation time was the same.

The value of the 100th best distance varied somewhat depending on the species tree, with smaller values corresponding to larger branch lengths in the species tree. This is due to larger branches meaning that there is less gene tree incongruence and therefore more observed and simulated trees will have the same unrooted topology as the species tree. The 100th best squared Euclidean distance varied from 16 to 266 for topologies using the trees in Table 1. Squared distances for splits were slightly higher, although we note that because n -taxon trees contribute $n - 3$ nontrivial splits each, the maximum distance using split counts is higher than that using topology counts. For splits, distances also tend to increase with more taxa or loci due to each tree contributing multiple splits.

To examine the effects of using estimated gene trees rather than known gene trees, 3 of the 14 species trees, one of each shape—caterpillar, pseudocaterpillar, and balanced—were chosen to repeat the simulation. The branch lengths chosen were based on those that performed best using known gene trees, with the idea that performance would be expected to be worse when using estimated gene trees (Fig. 3). The Seq-Gen program (Rambaut and Grassly, 1997) was used to generate DNA sequences of length 500 nucleotides using an *HKY* + Γ model with base frequencies of 0.3, 0.2, 0.2, and 0.3 for A, C, G, and T, respectively. The program PHYML (Guindon and Gascuel, 2003) was used to estimate the unrooted gene trees. Details of the settings are provided in the scripts provided in the online supplement.

2.2.2. Simulation of Eight-Taxon trees—Algorithm 3 is used to compute the posterior probability when eight-taxon trees were inferred. The ape library (Paradis *et al.*, 2004) and the sets library (Hornik and Meyer, 2009) from the R-package (Ihaka and Gentleman, 1996) are used for calculating the symmetric difference between the multisets of splits for the distance between the simulated data and the observed data. This process works to find the splits that are in the observed data but not in the simulated data and inversely to find the splits in the simulated data but not in the observed data. Figure 4 shows the eight-taxon trees used for simulation and their corresponding unrooted topologies.

2.2.3. Simulation of 16-taxon trees—For 16-taxon trees, we also used Algorithm 3, using only split counts. Because there are so many tree shapes for 16 taxa, we simulated 50 species trees using TreeSim (Stadler, 2011) under a pure birth model with birth parameter $\lambda = 1.0$, which corresponds to an average branch length of 1.0 coalescent units. This is a similar approach to other species tree inference papers that have summarized results over species trees from a pure birth model (Huang *et al.*, 2010). The posterior distribution estimated using ABC is summarized using a consensus tree of the best 100 trees using Algorithm 3, again using 100 loci and $J = 50,000$ simulated data sets. Using the consensus tree to summarize a posterior distribution of trees is fairly standard in Bayesian phylogenetics (Holder *et al.*, 2008). As before, the prior for the topology is uniform over all possible rootings (29 possible root locations corresponding to 16 pendant branches and 13 internal branches), and the prior for the branch lengths is exponential with rate 1.

2.3. Empirical example

To apply the ABC method to an empirical example, we used a data set with 310 nuclear genes found in 45 seed plants analyzed by Xi *et al.* (2014). The study compared coalescent methods (STAR and MP-EST) with concatenation, and found that the coalescent methods found that *Nuphar* clustered with *Amborella* rather than the ingroup taxa, whereas using concatenation, there was strong support for *Amborella* being an outgroup to the flowering plants including *Nuphar*.

The gene trees from Xi *et al.* (2014) have missing taxa for many loci. We therefore considered two subsets of taxa with the aim of having as many loci as possible. A subset of five taxa is used for which there were 275 loci, and another subset of eight taxa is taken which yields 224 nuclear genes.

To apply the ABC method for the angiosperm study, we apply algorithms 1 and 2 when there are five taxa, and algorithm 3 for eight taxa, using empirical gene tree topologies instead of simulated gene trees for step (1). The hypotheses of Xi *et al.* (2014) are used as priors — their first hypothesis is that *Amborella* clusters with *Nuphar*, and their second hypothesis is that *Nuphar* clusters with all other taxa (Figs. 5 and 6). Thus, the prior is restricted to two rooted topologies, which should make the ABC algorithm more efficient because half of the simulated data sets are expected to be simulated from a species tree with the correct topology (as opposed to 1/7 and 1/13 for the five- and eight-taxon simulations elsewhere in the paper).

3. Simulation Results

3.1. 5 taxa, known gene trees

Tables 1 and 2 summarize the posterior distributions using topology counts and split counts, respectively, for the five-taxon simulations with known gene trees. We report the proportion of times that the highest posterior probability tree (also called the MAP tree for maximum a posteriori) matched the species tree (proportion correct), the proportion of times that the correct species tree was in the 90% posterior credibility region (coverage probability), the number of trees in the 90% credibility region, and the average posterior probability over 50 replicates for the seven possible rooted species trees in the prior, $T_1, \dots, T_7$ (Fig. 2).

The most accurately inferred caterpillar species trees is $S_{c,2}$, which has two short internal branches and a relatively long branch incident to the root. It has the highest proportion correctly inferred at 94% (using either topology or split counts), the highest coverage probability at 100%, the highest average posterior probability at ~68% and ~60% using topologies and splits, respectively, and the smallest credibility region at fewer than 3 trees for 90% credibility for both topologies and splits. The second most accurately inferred caterpillar is $S_{c,1}$, which has all short branches. In fact, $S_{c,1}$ and $S_{c,2}$ are the only species trees in the five-taxon study that are in the unrooted anomaly zone, meaning that for these trees, the most probable unrooted gene tree has a different unrooted topology from that of the species tree (Degnan, 2013). In usual species tree inference problems, species trees with longer branches make inference easier (Huang *et al.*, 2010; Wu, 2012). We note that for the caterpillar topology, species trees with shorter branches had a higher probability of being

recovered correctly, and the tree with the longest branches (all branches being 1.0 coalescent units) had the worst performance.

Similarly to the caterpillar species trees, the pseudocaterpillar trees that were most accurately inferred had either all short branches, or short branches with a relatively long branch incident to the root (Tables 1,2). Also similar to the caterpillar species trees, the pseudocaterpillar that was most difficult to infer had the longest branches.

Among the balanced species trees, the most accurately inferred is $S_{b,2}$, since it has the highest proportion correct at 0.65, the highest coverage probability at 92%, and the highest average posterior probability at 29% for the correct tree; however, its credibility region of 5.26 is not the lowest desirable among the balanced species trees. The pattern between long and short branches compared to the caterpillar species tree seems reversed for the balanced tree: the tree with the longest branches was recovered the most frequently, which is more similar to the usual pattern for species tree inference problems.

The balanced species tree which is the most difficult to infer is $S_{b,5}$, because the proportion correct is as 0%, the coverage probability is very low at 2%, and the average posterior probability is very low at 7.14%. We note that for the balanced trees that were the least accurate, $S_{b,1}$ and $S_{b,5}$ still recovered some of the clade information accurately. The posterior probability for a particular clade is the sum of the posterior probabilities for all trees that have the given clade. Thus, for $S_{b,5}$ in Table 1, although it never recovered all three clades correctly in the MAP tree, the average posterior probability for the (a, b) clade was 97% (summing posterior probabilities for trees T_1 , T_2 , T_5 , T_6 , and T_7), and the average posterior probability for the (a, b, c) clade was 67% (trees T_1 , T_2 , and T_6) (Table 1). Thus, the difficulty was in recovering the (d, e) clade. Using PHYLIP (Felsenstein93) to compute the average rooted RF distance between the extended majority rule consensus tree and the species tree, we get 1.32 and 1.36 using topologies and splits, respectively for $S_{b,1}$ and 1.24 and 1.16 using topologies and splits, respectively for $S_{b,5}$. These RF distances range from 19.3% to 22.7% of the maximum RF distance of 6. Getting one clade wrong in the consensus tree would give an RF distance of 2 (33% of maximum).

Results using either topology counts or split counts as the summary statistic tended to be highly similar. The proportion of trees correctly inferred using splits rather than topology counts was a little lower for two of the caterpillar species trees ($S_{c,4}$, $S_{c,5}$) and a little higher for the balanced trees. Coverage probabilities were also very similar overall. Posterior probabilities for the correct tree tended to be slightly lower when using splits, but are highly correlated with posterior probabilities using topology counts (correlation > 0.99 ; Fig. 7). The average number of trees in the 90% credibility regions also tended to be slightly higher using splits but still very similar.

The similar performance using splits as a summary statistic suggests that using splits for larger trees might be reasonable. The advantage in using splits for larger trees is that the number of splits grows more slowly than the number of topologies, and so an ABC algorithm using splits rather than topology counts might face less of a problem in terms of dimensionality on larger trees.

We also investigated the effect of having significantly longer branches, as well as having a star-shaped species tree with all internal branches having length 0. We simulated for the caterpillar and balanced tree topologies with $x = y = z = 0, 0.1, 0.5, 1.0, 2.0$, and 3.0 using the same settings as before, with $J = 50,000$ and retaining the best 100 proposed species trees, with 50 replicates for each combination of topology and branch lengths.

When all branches have the same length, and are long in coalescent units, there is less variation in the gene trees. In this case, a large number of gene trees have the topology $((a, b), c, (d, e))$, which matches the unrooted topology of all of the species trees in the prior. If all gene trees have this topology, then there should be little or no information in the data to determine the correct rooted species tree. In this case, we might expect that the highest posterior probability tree would be equally probable to be any of the seven trees in the prior.

For the caterpillar topology, it is observed that for long branches, the proportion of times that the correct species tree is observed is highest when $x = y = z = 0.5$, and quickly goes to 0 when the branch length is 2 or more coalescent units. When the branch lengths are 0 (i.e., a star tree), the proportion of times that the caterpillar is inferred is higher than the value of $1/7$ that would be expected if all topologies were equally probable, with a proportion of 26% using topology counts (Fig. 8). This suggests that the prior is informative. Similarly, having 0% probability of recovering the correct tree for long branches suggests an informative prior. For these cases, the prior, although uniform for the topology, also includes a prior for branch lengths that is not well suited to the data.

Caterpillar species trees tend to have higher gene tree discordance than balanced species trees given similar branch lengths (Degnan and Salter, 2005), which could explain why the caterpillar is favored in the posterior when gene trees are purely random (the star species tree) and why caterpillars are under-estimated when the gene trees have no variation (long species tree branches). Consistent with this prediction, the proportion of times the balanced species tree is inferred is lower than would be expected for an uninformative prior for the star tree, and there is a bias in favor of balanced topologies when there is no variation in the gene tree topologies (Fig. 8). These examples also illustrate that, unlike the case of inferring rooted species trees from rooted gene trees, or unrooted species trees from unrooted gene trees, longer internal branches do not necessarily make the inferences easier. Inference of the rooted species tree from unrooted gene trees requires variation in the gene trees.

3.2. 5 taxa, gene trees estimated from DNA sequences

Tables 3 and 4 show the same measures of accuracy as Tables 1 and 2 and both display that the coverage probability contains the correct species in all the iterations for both species trees. Using the caterpillar species tree $S_{c,2}$, the highest posterior tree always matched the species tree and had an average posterior probability of 60.1% and 49.5% using topologies and splits, respectively. When species tree $S_{b,3}$ was used, the highest posterior probability tree matched the species tree only 35% and 20% of the time using topologies and splits, respectively. For these examples, inference of the caterpillar species tree was improved using estimated rather than known gene trees, while inference of the balanced species tree was somewhat worse for estimated gene trees. Although estimated rather than known gene trees typically make species tree inference more difficult (Huang *et al.*, 2010; Roch and Warnow,

2015), we speculate that biases in estimated gene trees (Huelsenbeck and Kirkpatrick, 1996) might lead to a gene tree distribution which favors some trees at the expense of others, and note that improved performance in the anomaly zone using estimated gene trees rather than known gene trees has been observed previously (Wang and Degnan, 2011).

A difference between the known gene tree versus estimated gene tree cases was that with estimated gene trees, there was much less correlation between posterior probabilities based on topologies versus splits. For both species trees used with estimated gene trees, $S_{c,2}$ and $S_{b,3}$, the correlation between the two posterior probabilities was not statistically significant ($p = 0.67$ and $p = 61$, respectively); however, only 10 iterations were used for the case of estimated gene trees due to the computation time involved.

3.3. Eight taxa, known gene trees

To calculate the posterior probabilities when the species tree had eight taxa, only split counts are used in the gene trees. As seen in the highest posterior probability tree never matched the species tree, but instead always had a (G, H) clade with all other clades being correct (Fig. 9). The RF distance was therefore always equal to 2.0. The posterior probability when the species tree is balanced trees is much flatter than it is for the caterpillar species tree, with the number of trees in the 90% credibility region being usually 10 out of 13 trees of the prior, but in the caterpillar tree, it is usually 6 out of 13 trees of the prior. The correct balanced tree was always obtained in all 50 iterations. For both trees, the the coverage probability was 1.0.

3.3.1. 16 taxa, known gene trees—For 16-taxon trees, the consensus trees of the 100 best species trees had an average RF distance of 3.8 (13.5% of maximum) from the species tree. Here an RF of 4 corresponds to getting two clades out 14 incorrect in the species tree. There were 5 out 50 trees recovered completely accurately (RF=0), and 46% of trees had an RF of 2 or lower (i.e., at most one clade is incorrect). If these unrooted species trees were rooted at random (assuming that the unrooted species tree is correct), then the average RF distance would be 5.9 (i.e., 21% of clades would be incorrect on average). Thus, the 50 16-taxon trees did significantly better than randomly rooting the trees ($p = 5.35 \times 10^{-7}$ based on a single sample t -test).

There is a clear relationship between the shape of the tree and accuracy, with trees that are more balanced tending to have lower RF distances to the true species tree (Fig. 10). To measure this effect, we used the Colless statistic, which measures the sum of differences between the number of left versus right (L_i versus R_i) descendants from internal nodes i (Colless, 1982):

$$I_c = \sum_{i \in \{\text{nodes}\}} |L_i - R_i|.$$

We plot the scaled RF distance, which divides the RF distance by $2n - 4$ against the Colless statistic. This scaling gives the proportion of clades that are incorrect in the estimated tree and is commonly used in measures of accuracy for species tree methods (e.g., Huang *et al.*,

2010). The correlation between the Colless statistic and the RF distance is 0.75 ($p = 3.7 \times 10^{-5}$).

These species trees were generated under a pure birth model with $\lambda = 1$. Different results might be obtained for other values of λ . For example, with larger values of λ , branch lengths tend to be shorter. This causes more gene tree incongruence, which might cause ABC to favor less balanced trees and might change the relationship between tree imbalance and accuracy.

4. *Amborella* example

The ABC method shows greater support for the *Amborella* and *Nuphar* clade than for the *Amborella*-as-outgroup hypothesis; however, the support is not overwhelming, with the posterior probability for this hypothesis being either 59% or 86% for the eight-taxon and five-taxon analyses, respectively. This lends some, but not overwhelming, support to the hypothesis of Xi et al. (2014) as opposed to the hypothesis supported by concatenation, using a subset of their gene trees. This is perhaps not surprising since we used the same gene tree topologies that lead to their conclusion and also using a coalescent method, although it is quite different from the methods used in their study.

Simmons and Gatesy (2015) argue that the anomaly zone, in which the most probable gene tree has a different topology from that of the species tree, does not apply to the rooting of the angiosperms because there is only one short branch leading to the *Amborella*-*Nuphar* clade, whereas the anomaly zone typically occurs when there are two consecutive branches on a path from the root to the tips (Degnan, 2013; Rosenberg, 2013). Although this is typical of species trees in the anomaly zone, a species tree with only one short branch can still be in the anomaly zone if the less basal branch can be indefinitely long if the more basal branch is sufficiently short (Degnan and Rosenberg, 2006). However, for this example, there is another reason for thinking that the anomaly zone is not a factor, which is that caterpillar gene tree shapes, which the *Amborella*-only outgroup hypothesis implies for the subset of species we analyzed, cannot be anomalous gene trees (Degnan and Rhodes, 2015).

Interestingly, if an extended majority rule (also called greedy) consensus tree is made of either the 275- or 224-loci datasets that we used, the resulting consensus tree agrees with the concatenation-based tree, which shows *Amborella* as the sole outgroup. In several simulation studies where concatenation and greedy consensus have disagreed with more coalescent-aware methods, concatenation and greedy consensus have both tended to support topologies that are anomalous gene trees—gene trees that are more probable under the multispecies coalescent than the gene tree with the same topology as that of the species tree—or trees that are more balanced than the species tree (Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007; Degnan et al., 2009; Liu et al., 2009; DeGiorgio and Degnan, 2010), and theoretical arguments suggest that concatenation can be attracted to a more balanced tree than the underlying species tree (Roch and Steel, 2015). In this example, however, concatenation and extended majority rule support a caterpillar tree, even though theoretical considerations prove that caterpillar gene trees can never be anomalous gene trees (Degnan and Rhodes, 2015). It has not been shown so far in theory or in simulation, to our

knowledge, that there are cases where the species tree has a non-caterpillar topology and either concatenation or extended majority rule consensus would theoretically tend to infer a caterpillar topology, so this makes the Xi *et al.* (2014) data set interesting.

Another consideration is the clade support for *Amborella* as an outgroup. Simmons and Gatesy (2015) point out that the *Amborella*-only outgroup occurs more often than any conflicting relationship. Under the multispecies coalescent model, any clade that occurs with more than 1/3 probability in the true gene trees is guaranteed to be in the species tree (Allman *et al.*, 2011a). Using the 275-locus five-taxon dataset, the proportion of loci for which *Amborella* is the only outgroup is $107/275 = 38.9\%$ (p-value = .055 or .030 for two- and one-sided tests, respectively), thus suggesting some evidence using coalescent considerations that the *Amborella*- outgroup hypothesis is correct. The five-taxon data set with more trees has weaker evidence, with $84/224=37.5\%$ of trees supporting the *Amborella*- hypothesis (p-value = .20 or .11 for two- and one-sided tests, respectively). Although the proportion of trees supporting the *Amborella*-only outgroup hypothesis is larger than for trees with the *Amborella*-*Nuphar* clade, these differences in proportion are also not significant. Simmons and Gatesy (2015) also criticize the data analysis that lead to the 310 gene trees of Xi *et al.* (2014), particularly in terms of rooting, but also in terms of the alignments used to reconstruct the gene trees. Simmons and Gatesy (2015) prefer the summary coalescent analysis of Wickett *et al.* (2014), which used the unrooted method ASTRAL (Mirarab *et al.*, 2014) and found support for the *Amborella*-only outgroup hypothesis. Our analysis is not intended to take sides in this debate; instead, we used the Xi *et al.* (2014) data to illustrate how the ABC approach can be used to estimate a root from unrooted gene trees, and find that this data set has the interesting property that a set of gene trees can lead to a naive consensus method returning a caterpillar tree while other methods can lead to a non-caterpillar estimate of the species tree.

5. Discussion

Our method is the first method we are aware of to explicitly use the coalescent model to estimate rooted species trees from unrooted gene trees; however, a version of the minimize deep coalescence method (MDC) was also developed to infer rooted species trees from unrooted gene trees (Yu *et al.*, 2011). When two or more lineages fail to coalesce in a population (i.e., incomplete lineage sorting occurs), this results in extra lineages that persist through speciation events going backward in time. The MDC score counts the minimum number of these extra lineages needed for a gene tree to occur in a particular species tree. MDC was originally developed based on the idea of inferring the rooted species tree that required the smallest number of extra lineages when reconciling the gene trees with the species tree. This was extended to unrooted gene trees by additionally minimizing over the root locations of the gene trees and can be used to return a rooted species tree using unrooted gene trees as input (Yu *et al.*, 2011). In other words, the MDC score contributed by a gene tree for a candidate species tree is computed as the minimum number of extra lineages where the minimum is taken over all possible rootings of the gene tree. By contrast, in the ABC method we have developed, we never infer the root of the gene trees.

The MDC criterion, although one of the first to be implemented to infer species trees from gene trees (Maddison and Knowles, 2006), was found to be statistically inconsistent in the rooted setting (i.e., using rooted gene trees as input) in the same year that its unrooted extension was published (Than and Rosenberg, 2011), meaning that it can be increasingly likely to return an incorrect species tree as the number of input gene trees goes to infinity, and has not been used as frequently since more accurate methods have been developed. We also note that PhyloNet can use unrooted MDC to return a rooted species tree even in the case of four taxa, although four-taxon gene tree topologies do not identify the rooted species tree under the multispecies coalescent (Allman *et al.*, 2011b). Although our focus has not been on comparing our ABC method to the unrooted version of MDC, we did a brief simulation with five taxa to check its performance using the three five-taxon shapes (caterpillar, pseudocaterpillar, and balanced) and either all branch lengths equal to 0.1 or all equal to 1.0. Each simulation used 50 iterations and 100 loci, as in the ABC simulations. PhyloNet version 3.6.0 was used with the command `Infer ST MDC UR` and all trees tied for the best MDC score were returned.

The MDC species trees showed a strong bias for the pseudocaterpillar shape, so that this was the most frequently returned shape regardless of the underlying species tree. Consequently, the pseudocaterpillar was either the best or tied for the best species tree for every iteration for the caterpillar and balanced species trees for both branch length settings. When the species tree had a pseudocaterpillar shape, the correct species tree was returned 85% and 100% of the time with branch lengths of 0.1 and 1.0, respectively. Strong shape biases have been known to exist for rooted MDC in previous research (Than and Rosenberg, 2011; DeGiorgio *et al.*, 2014), and more work would be needed to understand its behavior in the unrooted setting.

5.1. ABC and sufficient statistics

An important issue in ABC is that summary statistics used are often not sufficient statistics, in which case it is not guaranteed that the estimated posterior distribution converges to the true posterior distribution (Marjoram and Tavaré, 2006; Csilléry *et al.*, 2010). It is still common in ABC analyses to use summary statistics that lack sufficiency (Aeschbacher *et al.*, 2012). In our case, we note that split counts are not sufficient statistics. It has been shown that they identify the unrooted species tree (Allman *et al.*, 2016), and we are working on proofs of identifiability for the rooted species tree as well. For a summary statistic T to be a sufficient statistic, it should be the case that given two data sets $\mathbf{x}$, and $\mathbf{y}$, if $T(\mathbf{x}) = T(\mathbf{y})$, then any inference about the parameter should be the same for the two data sets (Casella and Berger, 2002). However, an example of two data sets each with three trees illustrates that split counts can be the same for two data sets with different input trees (Table 5).

For this example, the split counts for the two datasets are identical, but the original trees are different, and the likelihoods for the two data sets are different based on using unrooted gene tree probabilities listed in Allman *et al.* (2011b). Although split counts are not, therefore, sufficient statistics, they still contain information about the root. In the five-taxon examples, using split counts did essentially as well as topology counts, suggesting that there was very

little loss of information in using splits instead of topologies, in spite of the lack of sufficiency.

5.2. Computation time

The ABC approach for estimating the root location is fairly slow due to the large numbers of computations, but we find that it scales reasonably well with the number of taxa. We estimate that the average times needed to run Algorithm 3 for one species tree inference with five, eight, and 16 taxa were 26.4, 28.6, and 30.4 hours, respectively, suggesting a roughly linear dependence on the number of taxa. This suggests that the computation scales reasonably well with the number of taxa, although we fixed $J = 50,000$. Possibly with more taxa it would be better to increase J since there are fewer species trees simulated in the prior that match the true species tree. An alternative approach for inferring the root suggested by Allman *et al.* (2011b) is to use maximum likelihood of the unrooted gene trees. This method has not yet been implemented; however, likelihood calculations for rooted gene tree topologies scale exponentially in the number of taxa (Degnan and Salter, 2005; Wu, 2012; Rosenberg, 2007; Disanto and Rosenberg, 2016a,b), suggesting that this method will not scale well in the number of taxa. Furthermore, the only current method available for calculating unrooted gene tree probabilities is to sum gene tree probabilities over all possible root locations. Thus calculating an unrooted gene tree probability for a 16-taxon tree requires summing 29 16-taxon rooted gene tree probabilities. The advantage of the ABC approach is that it does not depend on calculating the likelihood, and simulating gene trees from larger species trees scales well with the number of taxa.

When estimated gene trees are used instead of known gene trees, the ABC approach requires estimating $G \cdot J$ gene trees, where G is the number of gene tree loci and J is the number of simulated data sets. This approach takes longer due to the estimation of the gene trees but scales linearly in the number of loci.

5.3. Conclusions

This study concludes that the root location for five-taxon species trees is in practice difficult to infer with 100 loci, but it is in some cases possible. More work needs to be done to see the effect of using larger numbers of loci. The ability of ABC to infer the root also seems to be very sensitive to the combination of the topology and branch lengths, and unfortunately it was difficult to generalize from our examples, where trees with short branches were easier to infer for the caterpillar and pseudocaterpillar but harder to infer for balanced tree shapes for five taxa.

Because the correlation between the average of the posterior probability for topology and the splits is very strong, there seems to be very little loss of information using split counts instead of topology counts for five taxa using known gene trees. Sequence data is not needed for the ABC method since gene trees are its only source of data. Despite shortcomings with the ABC method, larger sample sizes may increase the accuracy and decrease the variability of the estimates. The ABC method could be used with the at prior or a more informative prior. We used a prior that was at but assumed that a particular unrooted species tree was known. If the unrooted species tree also had uncertainty, then this could be reflected by

making the prior include more rooted trees. Another possibility is to consider that under typical birth-death processes, some unrooted trees are more probable than others when there are 6 or more taxa (Steel, 2012), and the prior could be based on this rather than making each labeled topology equally likely in the prior.

Since it is difficult to estimate the root location with certainty, a possible application is to instead rule out implausible locations for the root, and in this case a credible region for the root might be desirable. In the five-taxon cases, 90% credibility regions tended to have about 3–5 trees, meaning that about half of the possible root locations had very low posterior probability.

According to the results from the empirical data, the ABC method weakly supports the conclusion of Xi *et al.* (2014) that *Amborella* and *Nuphar* cluster together based on their gene tree topologies, but using very different methodology (although theirs is also coalescent-based). We note that their gene trees have been criticized, in particular for having problems with how they are rooted. However, our method uses only unrooted gene tree topologies and still found some support for their estimated species tree. We have used this data set not to make a definitive claim about the rooting of the angiosperms, but rather to illustrate how the ABC method could be used for this type of problem.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institutes of Health R01GM117590. We thank the University of New Mexico Center for Advanced Research Computing for computational resources used in this work. We especially thank Ryan Johnson at CARC for scripting help. We thank Zhenxiang Xi for sending the angiosperm gene trees used in their analysis.

References

- Aeschbacher S, Beaumont MA, Futschik A. A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics*. 2012; 192(3):1027–1047. [PubMed: 22960215]
- Allman ES, Degnan JH, Rhodes JA. Determining species tree topologies from clade probabilities under the coalescent. *J. Theor. Biol.* 2011a; 289:96–106. [PubMed: 21867714]
- Allman ES, Degnan JH, Rhodes JA. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *J. Math. Biol.* 2011b; 62:833–862. [PubMed: 20652704]
- Allman ES, Degnan JH, Rhodes JA. Species tree inference from gene splits by unrooted star methods. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2016; arXiv:1604.05364. doi: 10.1109/TCBB.2016.2604812
- Ané C, Larget B, Baum DA, Smith SD, Rokas A. Bayesian estimation of concordance factors. *Mol. Biol. Evol.* 2007; 24:412–426. [PubMed: 17095535]
- Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. *Genetics*. 2002; 162:2025–2035. [PubMed: 12524368]
- Boykin LM, Kubatko LS, Lowrey TK. Comparison of methods for rooting phylogenetic trees: A case study using orcuttieae (poaceae: Chloridoideae). *Mol. Phylogenet. Evol.* 2010; 54(3):687–700. [PubMed: 19931622]

- Buzbas EO. On the article titled “Estimating species trees using Approximate Bayesian Computation (Fan and Kubatko, *Molecular Phylogenetics and Evolution* 59: 354–363). *Mol. Phylogenet. Evolut.* 2012; 65(3):1014–1016.
- Casella, G., Berger, RL. *Statistical inference*. 2. Duxbury; Pacific Grove, CA: 2002.
- Chifman J, Kubatko L. Quartet inference from snp data under the coalescent model. *Bioinformatics*. 2014; 30(23):3317–3324. [PubMed: 25104814]
- Colless DH. *Phylogenetics, the theory and practice of phylogenetic syst. biol.s.* Syst. Zool. 1982; 31:100–104.
- Csilléry K, Blum MG, Gaggiotti OE, François O. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*. 2010; 25(7):410–418. [PubMed: 20488578]
- DeGiorgio M, Degnan JH. Fast and consistent estimation of species trees using supermatrix rooted triples. *Mol. Biol. Evol.* 2010; 27:552–569. [PubMed: 19833741]
- DeGiorgio M, Syring J, Eckert AJ, Liston A, Cronn R, Neale DB, Rosenberg NA. An empirical evaluation of two-stage species tree inference strategies using a multilocus dataset from north american pines. *BMC Evolut. Biol.* 2014; 14(1):67.
- Degnan JH. Anomalous unrooted gene trees. *Syst. Biol.* 2013; 62:574–590. [PubMed: 23576318]
- Degnan JH, Rhodes JA. There are no caterpillars in a wicked forest. *Theor. Popul. Biol.* 2015; 105:17–23. [PubMed: 26365816]
- Degnan JH, Rosenberg NA. Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2006; 2:762–768.
- Degnan JH, Salter LA. Gene tree distributions under the coalescent process. *Evolution*. 2005; 59:24–37. [PubMed: 15792224]
- Degnan JH, DeGiorgio M, Bryant D, Rosenberg NA. Properties of consensus methods for inferring species trees from gene trees. *Syst. Biol.* 2009; 58:35–54. [PubMed: 20525567]
- Disanto F, Rosenberg NA. Asymptotic properties of the number of matching coalescent histories for caterpillar-like families of species trees. *IEEE/ACM Trans. Comp. Biol. Bioinform.* 2016a; 13(5): 913–925.
- Disanto F, Rosenberg NA. Enumeration of ancestral configurations for matching gene trees and species trees. *arXiv preprint arXiv:1610.07549*. 2016b
- Ewing GB, Ebersberger I, Schmidt HA, Von Haeseler A. Rooted triple consensus and anomalous gene trees. *BMC Evolut. Biol.* 2008; 8(1):118.
- Fan HH, Kubatko LS. Estimating species trees using approximate Bayesian computation. *Mol. Phylogenet. Evolut.* 2011; 59(2):354–363.
- Gadagkar SR, Rosenberg MS, Kumar S. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J. Exp. Zool. Part B*. 2005; 304(1):64–74.
- Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 2003; 52(5):696–704. [PubMed: 14530136]
- Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Syst. Biol.* 2010; 59(3):307–321. [PubMed: 20525638]
- Holder MT, Sukumaran J, Lewis PO. A justification for reporting the majority-rule consensus tree in bayesian phylogenetics. *Systematic biology*. 2008; 57(5):814–821. [PubMed: 18853367]
- Hornik K, Meyer D. Generalized and customizable sets in R. *Journal of Statistical Software*. 2009; 31(2):1–27.
- Huang H, Knowles LL. What is the danger of the anomaly zone for empirical phylogenetics? *Syst. Biol.* 2009; 58:527–536. [PubMed: 20525606]
- Huang H, He Q, Kubatko LS, Knowles LL. Sources of error inherent in species-tree estimation: Impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Syst. Biol.* 2010; 59:573–583. [PubMed: 20833951]
- Huelsenbeck JP, Kirkpatrick M. Do phylogenetic methods produce trees with biased shapes? *Evolution*. 1996; 50:1418–1424. [PubMed: 28565707]
- Ihaka R, Gentleman R. R: a language for data analysis and graphics. *Journal of computational and graphical statistics*. 1996; 5(3):299–314.

- Joyce P, Marjoram P. Approximately sufficient statistics and Bayesian computation. *Stat. Appl. Genet. Mo. B.* 2008; 7(1):26.
- Kubatko LS, Degnan JH. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 2007; 56(1):17–24. [PubMed: 17366134]
- Kubatko LS, Carstens BC, Knowles LL. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics.* 2009; 25(7):971–973. [PubMed: 19211573]
- Kupczok A, Schmidt HA, von Haeseler A. Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. *Algorithm. Mol. Biol.* 2010; 5(1):1.
- Larget BR, Kotha SK, Dewey CN, Ané C. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics.* 2010; 26(22):2910–2911. [PubMed: 20861028]
- Liu L, Yu L. Estimating species trees from unrooted gene trees. *Syst. Biol.* 2011; 60(5):661–667. [PubMed: 21447481]
- Liu L, Yu L, Pearl DK, Edwards SV. Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* 2009; 58:468–477. [PubMed: 20525601]
- Liu L, Yu L, Edwards SV. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolut. Biol.* 2010a; 10(1):302.
- Liu L, Yu L, Pearl DK. Maximum tree: a consistent estimator of the species tree. *J. Math. Biol.* 2010b; 60(1):95–106. [PubMed: 19283383]
- Maddison WP, Knowles LL. Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 2006; 55(1):21–30. [PubMed: 16507521]
- Marjoram P, Tavaré S. Modern computational approaches for analysing molecular genetic variation data. *Nat. Rev. Genet.* 2006; 7(10):759–770. [PubMed: 16983372]
- Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. Astral: genome-scale coalescent-based species tree estimation. *Bioinformatics.* 2014; 30(17):i541–i548. [PubMed: 25161245]
- Mossel E, Roch S. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2010; 7(1):166–171. [PubMed: 20150678]
- Paradis E, Claude J, Strimmer K. Ape: analyses of phylogenetics and evolution in the R language. *Bioinformatics.* 2004; 20(2):289–290. [PubMed: 14734327]
- Rambaut A, Grassly NC. Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comp. Appl. Biosci.* 1997; 13:235–238. [PubMed: 9183526]
- Robinson D, Foulds L. Comparison of phylogenetic trees. *Math. Biosci.* 1981; 53:131–147.
- Roch S, Steel M. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor. Popul. Biol.* 2015; 100:56–62.
- Roch S, Warnow T. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Syst. Biol.* 2015; 4:663–676.
- Rosenberg NA. Counting coalescent histories. *J. Comput. Biol.* 2007; 14(3):360–377. [PubMed: 17563317]
- Rosenberg NA. Discordance of species trees with their most likely gene trees: a unifying principle. *Mol. Biol. Evolut.* 2013; 30:2709–2713.
- Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature.* 2013; 497(7449):327–331. [PubMed: 23657258]
- Simmons MP, Gatesy J. Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms. *Mol. Phylogenet. Evol.* 2015; 91:98–122. [PubMed: 26002829]
- Sisson SA, Fan Y, Tanaka MM. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA.* 2007; 104(6):1760–1765. [PubMed: 17264216]
- Solís-Lemus C, Ané C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.* 2016; 12(3):e1005896. [PubMed: 26950302]
- Stadler T. Simulating trees on a fixed number of extant species. *Syst. Biol.* 2011; 60:676–684. [PubMed: 21482552]
- Stamatakis A. RAxML-VI-HP: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 2006; 22(21):2688–2690. [PubMed: 16928733]

- Steel M. Root location in random trees: A polarity property of all sampling consistent phylogenetic models except one. *Mol. Phylogenet. Evol.* 2012; 65(1):345–348. [PubMed: 22772025]
- Than C, Ruths D, Nakhleh L. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics.* 2008; 9:322. [PubMed: 18662388]
- Than CV, Rosenberg NA. Consistency properties of species tree inference by minimizing deep coalescences. *J. Comput. Biol.* 2011; 18:1–15. [PubMed: 21210728]
- Wang Y, Degnan JH. Performance of matrix representation with parsimony for inferring species from gene trees. *Stat. Appl. Genet. Mol. Biol.* 2011; 10:1.
- Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. *P. Natl. Acad. Sci. USA.* 2014; 111(45):E4859–E4868.
- Wu Y. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution.* 2012; 66:763–775. [PubMed: 22380439]
- Xi Z, Liu L, Rest JS, Davis CC. Coalescent versus concatenation methods and the placement of amborella as sister to water lilies. *Syst. Biol.* 2014; 63(6):919–932. [PubMed: 25077515]
- 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):1543–1559. [PubMed: 22035329]
- Zhu S, Degnan JH, Goldstien SJ, Eldon B. Hybrid-lambda: simulation of multiple merger and kingman gene genealogies in species networks and species trees. *BMC Bioinformatics.* 2015; 16(1):292. [PubMed: 26373308]

Highlights

- A new method of species tree inference is proposed.
- The method uses unrooted gene tree topologies as input and returns a rooted species tree
- The method is based on Approximate Bayesian Computation
- The method is applied to the rooting of the angiosperms

Figure 1.

Example of computing distances using multisets. Suppose the top row of trees is the observed data and the bottom row is a simulated data set. We use notation such as 14|– to indicate the split 14|235678. The splits observed in the top row as a multiset (i.e., keeping track of how often each split occurs) are

$$S_{\text{obs}} = \{14|-, 14|-, 14|-, 23|-, 56|-, 56|-, 56|-, 78|-, 78|-, 78|-, 256|-, 378|-, 378|-, 1456|-, 1456|-\}.$$

The splits observed in the bottom row as a multiset are

$$S_{\text{sim}} = \{14|-, 14|-, 14|-, 56|-, 56|-, 56|-, 78|-, 78|-, 78|-, 124|-, 124|-, 378|-, 378|-, 378|-, 1456|-\}$$

The set difference $S_{\text{obs}} \setminus S_{\text{sim}}$ is the set of splits in S_{obs} that are not in S_{sim} , keeping track of how many times each split occurs in each set. Therefore $S_{\text{obs}} \setminus S_{\text{sim}} = \{23|-, 256|-, 1456|-\}$, which has size 3, and the set difference $S_{\text{obs}} \setminus S_{\text{sim}}$ is $\{124|-, 124|-, 378|-\}$, which also has size 3. For example, since 1456|– shows up twice in S_{obs} and once in S_{sim} , it occurs $2 - 1 = 1$ time in $S_{\text{obs}} \setminus S_{\text{sim}}$ but doesn't occur in $S_{\text{sim}} \setminus S_{\text{obs}}$. The total symmetric difference is therefore $3 + 3 = 6$.

Figure 2. Unrooted species tree and rooted trees in the prior. The species tree is assumed to have the unrooted topology of the top left tree. The prior for the five-taxon simulations is uniform on the seven possible rootings, $T_1, \dots, T_7$, of this unrooted tree.

Figure 3.
Species Tree Shapes and Branch Length to Simulate DNA Sequenaces

Figure 4.

Figure 5.

Two hypotheses from Xi *et al.* (2014) used as the prior for 8-taxa in the present study.

Numbers on the branches indicate the number of loci with the given clade in the empirical data. The tree on the right is the majority-rule consensus tree and also the tree inferred by concatenation.

Figure 6. Two hypotheses from Xi *et al.* (2014) used as the prior for 5-taxa in the present study. Numbers on the branches indicate the number of loci with the given clade in the empirical data. The tree on the right is the majority-rule consensus tree.

Figure 7. Correlation of average posterior probability using splits versus topologies for five-taxon trees. The points are for the case of known gene trees. The line $y = x$ is plotted so that points below the line indicate that the posterior probability was higher using topology counts rather than split counts as the summary statistic.

Figure 8. Proportion of times the species tree has the highest posterior probability when all internal branches are equal, $x = y = z$, for the caterpillar species tree (left) and balanced tree (right).

Figure 9.

Average posterior probabilities for clades in the most commonly estimated species trees for the 8-taxon simulation. (a) is the most frequently estimated species tree when the species tree had the caterpillar species tree (Fig. 4). (b) is the most frequently estimated species tree when the species tree has the balanced 8-taxon shape (Fig. 4). The { *G*, *H* } clade on the tree in (a) was the only clade not on the two model species trees).

Figure 10.
Relationship of tree balance to RF distance for the 16 taxon example.

Table 1

Average posterior probabilities using topology counts for five-taxon trees. Average posterior probabilities for the correct species tree topology are in bold.

Species tree	Species tree, branch lengths (x, y, z)	Prop. Correct (%)	Coverage probability (%)	Average of the posterior probability (%)							# of trees in 90% CR.
				T ₁	T ₂	T ₃	T ₄	T ₅	T ₆	T ₇	
Caterpillar	S _{c,1} , (0.1,0.1,0.1)	76	96	53.62	23.96	2.40	2.60	7.26	7.46	2.7	3.04
	S _{c,2} , (0.1,0.1,1.0)	94	100	67.66	20.8	0.30	1.64	4.04	4.70	0.86	2.40
	S _{c,3} , (1.0,0.1,0.1)	50	96	40.10	34.00	0.56	0.56	15.48	3.94	5.36	3.32
	S _{c,4} , (0.1,1.0,0.1)	47	100	36.06	33.00	0.28	0.36	1.30	28.86	0.14	3.06
	S _{c,5} , (1.0,1.0,1.0)	36	94	25.34	25.32	2.10	2.82	9.46	28.08	6.88	4.14
Pseudo-caterpillar	S _{p,1} , (0.1, 0.1, 0.1)	92	96	7.74	8.86	8.20	7.78	56.06	5.98	5.38	3.66
	S _{p,2} , (1.0, 1.0, 1.0)	9	54	16.32	16.18	12.12	13.78	12.46	14.96	14.18	4.40
	S _{p,3} , (0.1, 0.1, 1.0)	94	100	4.14	4.34	3.32	2.66	79.54	3.38	2.62	2.14
	S _{p,4} , (0.1, 1.0, 0.1)	52	96	2.46	2.84	24.22	20.90	31.74	12.80	5.02	4.22
Balanced	S _{b,1} , (0.1,0.1,0.1)	2	50	30.36	28.18	5.54	4.80	16.46	10.82	3.84	3.76
	S _{b,2} , (1.0,1.0,1.0)	65	92	15.94	14.82	8.52	8.08	13.02	28.74	10.88	5.26
	S _{b,3} , (0.1,0.1,1.0)	32	84	11.16	9.70	12.30	15.20	23.22	24.16	4.26	4.46
	S _{b,4} , (0.1,1.0,0.1)	26	88	34.60	36.06	0.54	0.38	1.28	27.02	0.12	2.98
	S _{b,5} , (1.0,0.1,0.1)	0	24	29.00	30.68	0.80	1.82	21.84	7.14	8.72	3.76

Table 2

Posterior using splits for five taxa. Posterior probabilities for the correct species tree topology are in bold.

Type of species tree	Species tree, branch lengths (x, y, z)	Prop. Correct (%)	Coverage probability (%)	Average of the posterior probability (%)							# of trees in 90% CR.
				T ₁	T ₂	T ₃	T ₄	T ₅	T ₆	T ₇	
Caterpillar	S _{c1} , (0.1,0.1,0.1)	76	94	47.96	24.74	4.46	4.04	10.28	6.10	2.42	3.46
	S _{c2} , (0.1,0.1,1.0)	94	100	59.60	27.64	0.96	2.00	5.20	3.64	0.96	2.72
	S _{c3} , (1.0,0.1,0.1)	52	94	37.70	32.32	0.66	0.48	19.08	3.68	6.08	3.34
	S _{c4} , (0.1,1.0,0.1)	35	100	35.30	33.28	0.80	0.46	2.16	27.80	0.20	3.06
	S _{c5} , (1.0,1.0,1.0)	33	94	24.46	24.62	2.04	2.54	11.34	26.88	8.12	4.20
Pseudo-caterpillar	S _{p1} , (0.1, 0.1, 0.1)	76	100	9.74	10.60	10.52	10.88	48.8	4.72	4.70	4.14
	S _{p2} , (1.0, 1.0, 1.0)	10	56	16.34	15.6	11.94	13.38	13.04	15.84	13.84	4.52
	S _{p3} , (0.1, 0.1, 1.0)	96	100	7.00	7.20	6.12	5.38	66.96	3.64	3.56	3.52
	S _{p4} , (0.1, 1.0, 0.1)	56	94	2.44	2.52	24.3	23.24	30.54	12.38	4.5	4.22
Balanced	S _{b1} , (0.1,0.1,0.1)	0	34	27.78	26.60	7.06	7.60	19.00	8.86	3.10	4.14
	S _{b2} , (1.0,1.0,1.0)	72	92	15.76	15.26	9.08	8.08	12.06	28.84	10.92	5.40
	S _{b3} , (0.1,0.1,1.0)	40	84	9.54	9.36	14.60	15.90	23.16	23.62	3.82	4.62
	S _{b4} , (0.1,1.0,0.1)	36	94	33.68	33.22	1.10	1.14	2.12	28.58	0.16	3.16
	S _{b5} , (1.0,0.1,0.1)	0	22	28.28	29.14	0.80	1.58	24.32	6.34	9.54	3.86

Table 3

Posterior probabilities for five-taxon species trees using estimated gene trees and topology counts.

Type of tree	Branch lengths (x, y, z)	Prop. correct (%)	Coverage probability (%)	Average of the posterior probability							# of trees in 90% CR.
				T_1	T_2	T_3	T_4	T_5	T_6	T_7	
caterpillar	(0.1,0.1,1.0)	100	100	60.1	21.5	1.8	1.9	7.2	5.0	2.5	3.3
pseudocaterpillar	(0.1,0.1,1.0)	100	100	8.1	6.3	5.9	7.8	61.3	5.3	5.3	4.9
balanced	(0.1,0.1,1.0)	35	100	6.4	7.0	21.4	20.4	19.6	20.8	4.4	5.4

Table 4
Posterior probabilities for five-taxon species trees using estimated gene trees and split counts

Species tree	branch lengths (x, y, z)	Prop. correct (%)	Coverage probability (%)	Average of the posterior probability (%)							# of trees in 90% CR.
				T ₁	T ₂	T ₃	T ₄	T ₅	T ₆	T ₇	
caterpillar	(0.1,0.1,1.0)	100	100	49.5	21.2	4.9	4.9	11.3	5.1	2.9	4.4
pseudocaterpillar	(0.1,0.1,1.0)	100	100	11.0	9.1	8.6	9.6	51.1	5.1	5.5	5.2
balanced	(0.1,0.1,1.0)	20	100	6.9	7.5	19.7	19.6	22.2	19.8	4.3	5.3

Table 5

Example to illustrate that split counts are not sufficient statistics

Dataset	tree	splits
1	t_1	$ABC DE, BC ADE$
1	t_2	$ABD CE, AB CDE$
1	t_3	$BCD AE, BD ACE$
2	t_4	$ABC DE, AB CDE$
2	t_5	$ABD CE, BD ACE$
2	t_6	$BCD AE, BC ADE$