SODA: Multi-locus species delimitation using quartet frequencies

Maryam Rabiee[1] and Siavash Mirarab[2]

[1]Computer Science and Engineering, University of California, San Diego, US
[2]Electrical and Computer Engineering, University of California, San Diego, US.

Abstract

Species delimitation, the process of deciding how to group a set of organisms into units called species, is one of the most challenging problems in evolutionary computational biology. While many methods exist for species delimitation, most based on the coalescent theory, few are scalable to very large datasets and those methods that scale tend to be not very accurate. Species delimitation is closely related to species tree inference from discordant gene trees, a problem that has enjoyed rapid advances in recent years. A major advance has been the surprising accuracy and scalability of methods that rely on breaking gene trees into quartets of leaves. In this paper, we build on this success and propose a new method called SODA for species delimitation. We test SODA in extensive simulations and show that it can easily scale to very large datasets while maintaining high accuracy.

keywords. Species delimitation, Phylogenetics, Multi-species coalescence

1 Introduction

Evolution results in diversity across species and diversity within the same species, in ways that can make it difficult to distinguish species. Definitions of what constitutes a species are varied (Coyne and Orr, 2004) and subject to debate. Nevertheless, many biological analyses depend on our ability to define and detect species. Assigning groups of organisms into units called species, a process called species delimitation, is thus necessary but remains challenging (Carstens *et al.*, 2013). Among varied species concepts, the most commonly used for Eukaryotes is the notion that individuals within a species should be able to mate and reproduce viable off-springs.

A wide range of species delimitation methods exist. More traditional methods simply relied on the mean divergence between sequences (e.g., Hebert *et al.*, 2004; Puillandre *et al.*, 2012) or patterns of phylogenetic branch length (Zhang *et al.*, 2013; Fujisawa and Barraclough, 2013; Esselstyn *et al.*, 2012) in marker genes or concatenation of several markers (e.g. Pons *et al.*, 2006). Due to limitations of marker genes (Hudson and Coyne, 2002), many approaches to species delimitation have moved to using multi-locus data that allow modeling coalescence within and across species (Knowles and Carstens, 2007; Yang and Rannala, 2010; O'Meara, 2010), not to mention more complex processes such as gene flow (e.g., Leaché *et al.*, 2019). Modeling coalescent allows methods to account for the fact that across the genome, different loci can have different evolutionary histories, both in topology and branch length (Maddison, 1997). Species delimitation is often studied using the Multi-species Coalescent (MSC) model (Pamilo and Nei, 1988; Rannala and Yang, 2003). In this model, individuals of the same species have no structure within the species and thus their alleles coalesce completely at random. Coalescence is allowed to cross boundaries of a species, resulting in deep coalescence events that can produce gene tree discordance due to Incomplete Lineage Sorting (ILS). In this context, given a set of sampled individuals, delimitation essentially requires inferring gene trees, one per locus, and detecting which delimitation is most consistent with patterns of coalescence observed in the gene trees.

Many methods for species delimitation under the MSC model exist, but they tend to suffer from one of two limitations. The most accurate set of methods are the Bayesian MCMC methods that infer gene trees, (optionally) species trees, and species boundaries (e.g., BPP (Yang and Rannala, 2010, 2014a), ABC (Camargo *et al.*, 2012), and STACEY (Jones, 2017)). Other Bayesian methods use biallelic sites (Leaché *et al.*, 2014), incorporate morphological data (Solís-Lemus *et al.*, 2015), or use structure (Huelsenbeck *et al.*, 2011). These methods, however, are typically slow and cannot handle even moderate numbers of samples (Fujisawa and Barraclough, 2013; Xu and Yang, 2016). A second class of methods (e.g., SpedeSTEM (Ence and Carstens, 2011)) rely on a three-step approach: first infer gene trees, then, date gene trees so that they all become ultrametric (i.e., have a unique root to tip distance), and finally, use ML calculation of alternative delimitations under the MSC model to decide species boundaries. These methods have been less accurate than Bayesian methods and their reliance on ultrametric trees make them hard to use for datasets where rates of evolution change substantially across the tree (Camargo *et al.*, 2012). Yet other methods (e.g., O'Meara, 2010; Zhang and Cui, 2010) rely only on input gene tree topologies, as we will do.

In this paper, we introduce a new species delimitation approach called SODA that builds on the success of our species tree inference tool ASTRAL (Mirarab *et al.*, 2014a; Mirarab and Warnow, 2015; Zhang *et al.*, 2018). As a statistically consistent method, ASTRAL infer a species tree from a collection of gene tree topologies (ignoring branch lengths) based on the principle that the most frequent unrooted topology for each quartet of species is expected to match the species tree (Allman *et al.*, 2011). Thanks to its accuracy and scalability, ASTRAL has been widely adopted for species tree inference. In a recent paper that extended ASTRAL to multi-individual data, we observed that if species boundaries are ignored, ASTRAL most often recovers individuals of the same species as monophyletic (Rabiee *et al.*, 2019). This result suggests a species delimitation method: Infer an ASTRAL tree with all individuals and use patterns of quartet trees mapped onto that species tree to decide where coalescence is completely random and where it is not; these boundaries readily define species. By relying on quartet frequencies and ASTRAL machinery, SODA is able to handle very large datasets with short running times. We describe SODA in detail and then perform two sets of extensive simulation studies to evaluate its accuracy and scalability.

2 Methods

2.1 Problem definition

We take a two-step approach to species delimitation where we assume unrooted gene tree *topologies* are already inferred. We then define the following problem.

Definition 1 (Species partitioning problem)**.** Given a set of unrooted gene trees $\mathcal{G}$ on a leafset $\mathcal{L} = \{l_1 \dots l_m\}$, find the maximum rank partition of $\mathcal{L}$ into $\mathcal{S} = \{S_1 \dots S_n\}$ such that for each i, the gene tree topologies restricted to S_i do not deviate from the free coalescence process.

The solution implies a mapping $m : \mathcal{L} \to \{1 ... n\}$ where $m(i)$ gives a species index for $i \in \mathcal{L}$. It differs from the formulation of Zhang and Cui (2010) in that we do not assume any known mapping or an explicit cost function. This formulation makes several assumptions. It ignores the population structure within the species and assumes that all individuals of the same species evolve according to the neutral Wright-Fisher model. Given this assumption, the distribution of gene trees within a

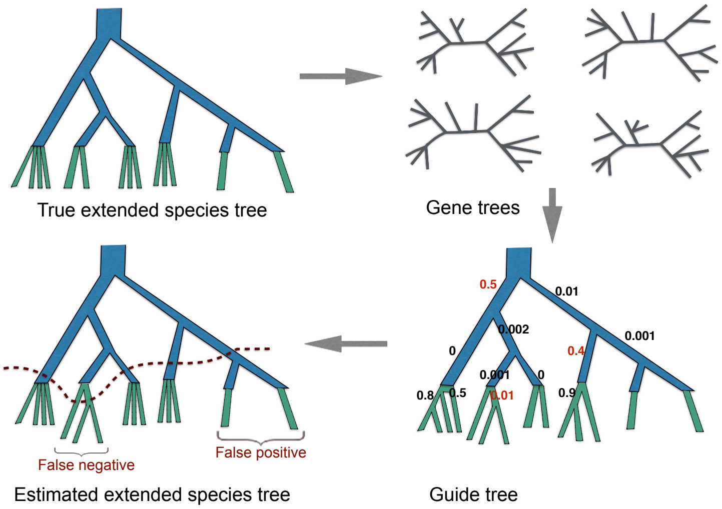

Figure 1: True extended species tree generates gene trees under the MSC model. From unrooted gene tree topologies (inferred from sequence data), SODA first estimates a guide tree using ASTRAL and then test the null hypothesis that each branch has length zero length, obtaining a p-value (bottom right). p-values may result in FP or FN rejection or retention of the null (red p-values). SODA then contracts some branches where the null is retained, while keeping others (e.g., those with p-value 0.4 and 0.5), in a way that ensures the resulting tree can be an extended species tree (bottom left). The inferred extended species tree can be cut at branches above the terminal branches to define species. The result could include both false positives and false negative delimitations. However, we note that some errors in p-values (0.5 and 0.4) are tolerated.

species should follow the Kingman (1982) coalescent process, motivating our problem formulation. It also assumes lineages from two species do not coalesce before their species separation, and thus, ignores gene flow across species. Finally, it ignores errors in gene trees estimated from data.

The branch lengths of gene trees can be modeled as a function of two processes: coalescent of lineages and changes in the mutation rate (Rannala and Yang, 2003). Dealing with these two processes simultaneously is a difficult computational challenge, motivating methods such as SpedeSTEM to take ultrametric (e.g., dated) gene trees as input and forcing Bayesian methods such as BPP to assume parametric rate models. In our work, we are after a fast delimitation method that can be applied to inferred non-ultrametric gene trees directly. To avoid complications of rate variations across lineages, we limit ourselves to gene tree topologies, a goal helped by the fact that the distribution of gene tree topologies under the MSC model is known (Degnan and Salter, 2005).

Using gene tree topologies, however, has a limitation. Examining the distribution of tree topologies requires at least three lineages. Thus, two species each with an individual sampled and one species with two individuals cannot be distinguished by topology alone. This forces us to assume that in the correct delimitation, each species has more than one individual sampled (i.e., $\forall_i |S_i| \geq 2$).

2.2 SODA Algorithm

We approach the species partitioning problem using the MSC model. A central concept in MSC is the "extended species tree", as defined by Allman *et al.* (2011). Let T^* be the true species tree on the leafset $\mathcal{S}$. The extended species tree $\mathcal{T}$ is a rooted tree labeled by $\mathcal{L}$, built by adding to each leaf of T^* all individuals corresponding to that species as a polytomy (Fig. 1); i.e., for leaf $s \in \mathcal{S}$ of T, add a child for every $r \in m^{-1}(s) \subset \mathcal{L}$.

Algorithm 1 - SODA Algorithm

```
function SODA($\mathcal{G}$, $T$,$\alpha$)
    if guide tree $T$ is not given then
        $T \leftarrow$ run ASTRAL on $\mathcal{G}$
    for internal branch $e$ of $T$ do
        $p(e) \leftarrow p$-value of the null hypothesis that length$(e) = 0$
    Root $T$ on the $\arg\min_e p(e)$
    ContractToExtendedSpeciesTre($T$,$\alpha$)
    partition $\mathcal{L}$ by cutting all (keep) internal edges of $T$
function ContractToExtendedSpeciesTree($T$,$\alpha$)
    for $e \in$ internal edges of $T$ do
        if $p(e) \leq \alpha$ then
            Mark $e$, sister of $e$, and all ancestors of $e$ as “keep”
    for $e \in$ internal edges of $T$ not marked “keep” do
        Contract $e$
```

Our species delimitation method, which we name Species bOundry Delimitation using Astral (SODA), is shown in Algorithm 1. Its inputs are a set of gene tree topologies and a confidence level α. SODA first infers (or takes as input) a guide tree T, assumed to be a resolution of the true extended species tree $\mathcal{T}$, then contract *some of the* branches of T that seem to have length zero given confidence level α, to get an estimated extended species tree $\hat{\mathcal{T}}$, and finally, cuts internal branches of the $\hat{\mathcal{T}}$ to cluster species (Fig. 1). Below, we describe each step in more details.

Guide Tree: SODA needs a (potentially unrooted) guide tree T on the leafset $\mathcal{L}$. If not provided by the user, we infer the guide tree by running ASTRAL-III on $\mathcal{G}$. We assume that T is a resolution of the extended species tree; thus, the accuracy of this guide tree is important.

Polytomy Test: For each branch of T, we need to test the null hypothesis that it has zero length in the coalescent units. If we cannot reject this null hypothesis, the branch can be collapsed to obtain a polytomy, helping us to obtain $\hat{\mathcal{T}}$. We use a recent test proposed by Sayyari and Mirarab (2018) that relies on a classic result: Under the MSC model, across gene trees, the frequencies of the three resolutions for each quartet around a given branch in the species tree are equal if and only if that branch has length zero (Pamilo and Nei, 1988; Allman *et al.*, 2011). To achieve scalability, this test treats branches independently (based on the locality assumption of Sayyari and Mirarab (2016)) and also treats quartets independently. The test produces one p-value per *internal* branch.

Rooting T: We need to root T (if not rooted) such that each species becomes monophyletic. We simply root T at the edge with the minimum p-value. Note that our goal is not to find the correct root because we do not need the correct rooting in the next steps. We only need the tree to be rooted on *any* internal branch of the extended species tree. The highest statistical confidence for having a positive length is achieved by the branch with the lowest p-value; thus we can root here.

Infer extended species tree. To obtain $\hat{\mathcal{T}}$, we contract some of the branches of T where a zero length null hypothesis cannot be rejected. When the null hypothesis is rejected for a branch e, we marked e as being part of $\hat{\mathcal{T}}$ (i.e., *keep* in Alg. 1). To ensure that we can get a valid extended species tree (with monophyletic species), we need polytomies to form only above terminal branches. Thus, in addition, we mark the sister edge of e and all its ancestor edges as belonging to $\hat{\mathcal{T}}$.

Partitioning: Given $\hat{\mathcal{T}}$, species partitioning is obtained by cutting internal branches of $\hat{\mathcal{T}}$.

The accuracy of the algorithm, in addition to assumptions made by our problem formulation, depends on the accuracy of the statistical test. We can formalize this notion in three claims (all proofs are straightforward and omitted here).

Claim 1. *Assuming (i) gene trees $\mathcal{G}$ are generated under the MSC model on an extended species tree $\mathcal{T}$, (ii) the guide tree T (e.g., ASTRAL tree) is a resolution of $\mathcal{T}$, and (iii) the hypothesis testing has no false positive (FP) or false negative (FN) errors, the SODA algorithm returns the correct extended species tree ($\hat{\mathcal{T}}$=$\mathcal{T}$) and hence the correct delimitation under the MSC model.*

This claim provides a reassuring positive result, but only under strong assumptions, most notably, that the test is perfect. However, errors in the test *can* lead to errors.

Claim 2. *Given a guide tree T that resolves the $\mathcal{T}$, SODA incorrectly divides a species S into multiple species (i.e., a false negative error) if and only if the zero length hypothesis testing results in an FP error for one of the branches under the clade defined by S on T.*

Thus, FPs in the hypothesis test always result in division of a species; however, an FP does not always divide S into exactly two parts. For example, if T has a caterpillar topology on S, an FP on the branch above the cherry leads to each remaining individual being marked as a species.

Claim 3. *Given a guide tree T that resolves $\mathcal{T}$, SODA incorrectly combines individuals from two species S_1 and S_2 into one species (a false positive error) under one of these two conditions.*

- *S_1 and S_2 each have one sampled individuals and form a cherry.*
- *The hypothesis testing has an FN error for all branches of $\mathcal{T}$ below the LCA of S_1 and S_2. This condition requires FN errors for two or more branches if neither species is a singleton.*

To combine two species incorrectly, we must fail to correctly mark all branches below their LCA; else, one of the branches below the LCA would be cut, which would prevent the FP. Thus, our approach is tolerant of some FN errors in the hypothesis test (e.g., p-value 0.4 in Fig. 1).

3 Simulation setup

3.1 Datasets

We used two simulated datasets, both generated using Simphy (Mallo *et al.*, 2016). One dataset is large and allows us to evaluate SODA on datasets where other methods cannot run whereas the other dataset is small and enables us to compare our method to slower Bayesian methods.

Large dataset: We reuse a 201-species "homogeneous" dataset that we have previously simulated (Rabiee *et al.*, 2019) using SimPhy to generate species trees based on the birth/death model and gene trees under the MSC model. Each species includes 5 individuals except the singleton outgroup (1001 in total). We have three model conditions (50 replicates each) with medium, high, or very high levels of ILS, with maximum tree height set to 2M, 1M, and 0.5M generations, respectively. The mean quartet score of true gene trees versus the species tree are 0.78, 0.62 and 0.50 for these model conditions (Fig. S5); the proportions of branches in the true extended species tree missing from gene trees are 0.40, 0.57 and 0.77. Each replicate has 1000 genes, which we down-sample randomly to 500, 200, and 100. To evolve nucleotide sequences down the gene trees, we use INDELible (Fletcher and Yang, 2009) with the GTR+Γ model of sequence evolution with randomly sampled sequence lengths from a LogNormal distribution (empirical mean=721). The average gene tree error for the three model condition is 0.25, 0.31 and 0.42 with large variance (Fig. S5).

Small dataset: We simulate a new dataset using SimPhy with 20 replicates, each only 4 species, 10 individuals per species, and 1000 genes per replicate (commands shown in Appendix A). The tree height is set to 200,000 generations, the population size is drawn uniformly between 10,000 and 500,000, and species trees are generated using birth-only model with rate=0.00001. These settings lead to high level of ILS, capturing a scenario close to where species delimitation is of relevance; the quartet score of true gene trees versus the true species tree is 0.76 and 65% of extended species tree branches are on average missing from true gene trees. We deviate from ultrametricity by drawing rate multipliers for species and genes from Gamma distributions, with LogNormal priors on parameters of Gamma (Table S1). We simulate 1000bp alignments on each gene tree using INDELible (Fletcher and Yang, 2009) and estimate gene trees using FastTree (Price *et al.*, 2010). The average gene tree error (normalized RF distance between true and estimated gene trees) is 43%. Gene tree distance between pairs of individuals of each species had a wide range in our simulations, falling between 10^{-5} and 10^{-2} mutations per site in most cases (Fig. S2).

3.2 Measures of accuracy

We evaluate accuracy using two measures.

ROC. Each pair of individuals is categorized depending on whether they are correctly grouped together (TP), correctly not grouped together (TN), incorrectly grouped together (FP), or incorrectly not grouped together (FN). We then show Recall = TP/(TP+FN) and FPR = FP/(TN+FP) on an Receiver Operating Characteristic (ROC) curve as we change α of SODA.

Adjusted Rand Index is a similarity measure between two partitions of a set also based on pairwise comparisons. We report ARI between the true species partition and the partition estimated by each method. The Rand (1971) index is $\frac{TP+TN}{TP+TN+FP+FN}$. The adjusted rand index (ARI) adjusts RI for expected similarity of pairs according to a generalized hyper-geometric distribution that controls for the number of objects and classes (Hubert and Arabie, 1985). ARI equals 1 only for the correct partition and is close to 0 for a random partition.

3.3 Methods compared

BPP. We compare SODA to the widely used method Bayesian Phylogenetics and Phylogeography (BPP). BPP uses MCMC for inferring species boundaries directly from sequence alignments by sampling gene trees and other model parameters (e.g., rates) under the MSC model. We use BPP 4.1.4 and took advantage of its multi-thread version to be able to run it with up to 1000 genes. Provided with all 40 individuals we sampled as 40 separate populations, BPP could not run to completion in 36 hours, perhaps because the set of possible delimitations was too large. To be able to test BPP, we randomly sample 4 individuals per species, and designate each as its own population. We use a uniform prior across all possible partitions on the resulting $4 \times 4 = 16$ populations. BPP calculates the posterior probability for each partition, and we use the delimitation with the highest posterior probability to measure the accuracy of the method.

We explore settings of BPP as follows. The total number of MCMC iterations is set to 208000 with the first 8000 discarded as burnin. We also run BPP with twice the number of iterations in one experiment with 500 genes to ensure convergence is not an issue. For the required species tree (similar to guide for SODA), we run BPP in two ways. By default, we provide BPP with the ASTRAL species tree (just as we do for SODA). The guide trees are rooted to match the true species tree. However, BPP is able to jointly infer species trees and species delimitation based on sequence alignments; we also run BPP with this co-estimation setting. This setting makes BPP 2× slower (even though we only have four species), making it impractical for an extensive study.

The priors for the inverse gamma distribution parameters, (α, β), were chosen to be (1.525,0.0001), (1.525,0.001) and (1.525,0.01) for population size θ_s and twice the β for τ_s. An example of a control file used for running BPP is given in Figure S1 for the full set of parameters.

SODA. We implemented SODA in python using Dendropy (Sukumaran and Holder, 2010). We vary α between 0.005 and 0.5 but designate $\alpha = 0.05$ as default. We infer the guide tree using ASTRAL-III on all 1000 genes. To control the impact of the guide tree, we also create the true extended species tree and resolve its polytomies randomly and use this guide tree as input to SODA.

4 Results

4.1 Large simulated dataset

On the large dataset with 1001 individuals, SODA takes no more than 35 minutes (Table S2) and is highly accurate (Fig. 2). Given 1000 genes, default SODA ($\alpha = 0.05$) is able to recover, on average, 183, 186, and 189 out of the 201 species entirely correctly for the three model conditions in decreasing order of ILS (Fig. 2a). The total number of species estimated by SODA ranges between 183 and 220 across all replicates with mean=208, which slightly over-estimates the correct number. The number of detected species increases as α increases; however, the number of correct species does not always increase. Reducing the number of genes reduces the number of correctly estimated species (down to 158 with 100 genes, very high ILS); however, it does not change the total number of species dramatically (for default α).

Examining pairs of individuals, we observe very high accuracy. With $\alpha = 0.05$, the ARI ranged between 0.95 and 0.97 for our three conditions given 1000 genes and between 0.87 and 0.92 when given as few as 100 genes (Fig. 2b). Reducing α to 0.005 or increasing it to 0.1 can reduce or increase ARI slightly; however, increasing α beyond 0.1 can quickly lead to substantial reductions in ARI (Fig. 2b). The best choice of α is always between 0.01 and 0.1, but 0.05 is never far from optimal, motivating us to use it as default (in addition to the tradition of using $\alpha = 0.05$ in statistical tests). Since our simulated replicates are very heterogeneous in terms of gene tree estimation error, we can also examine the impact of mean gene tree error on the accuracy of SODA. Interestingly, except for an outlier replicate with very high gene tree error and low ARI < 0.9, we do not detect a strong correlation between gene tree error and accuracy (Fig. S5).

The trade-off between precision and recall with different choices of α can be examined using the ROC curve (Fig. 2c). With $\alpha \leq 0.05$, recall is always 96% or higher, and is often close to 100% with $\alpha \leq 0.01$. The FPR, however, is strongly impacted by the number of genes. For example, with default α, FPR is never more than 0.03% with 1000 genes but increases to 0.98% with 100 genes. Reducing α reduces FPR, but for $\alpha > 0.1$, we observe little gain in FPR and a precipitous decline in recall as α increases. Thus, as observed earlier, $\alpha > 0.1$ does not seem advisable. Beyond the default value, a choice of $\alpha = 0.01$ seem desirable if more FP combinations can be tolerated. ROC curves also reveal interesting patterns in terms of the impact of ILS on the accuracy of SODA. For $\alpha = 0.05$ and a fixed number of genes, increasing ILS increases FPR (combining species) but does not substantially impact recall. Finally, using a random resolution of the true extended species tree as the guide tree has only a small positive impact on the accuracy (Fig. S4).

4.2 Small simulated dataset

On the small dataset with four species, SODA (default) has 98% recall with both 500 and 1000 genes (Fig. 3a). Increasing the number of genes mostly reduce FPR, from 14% with 500 genes to

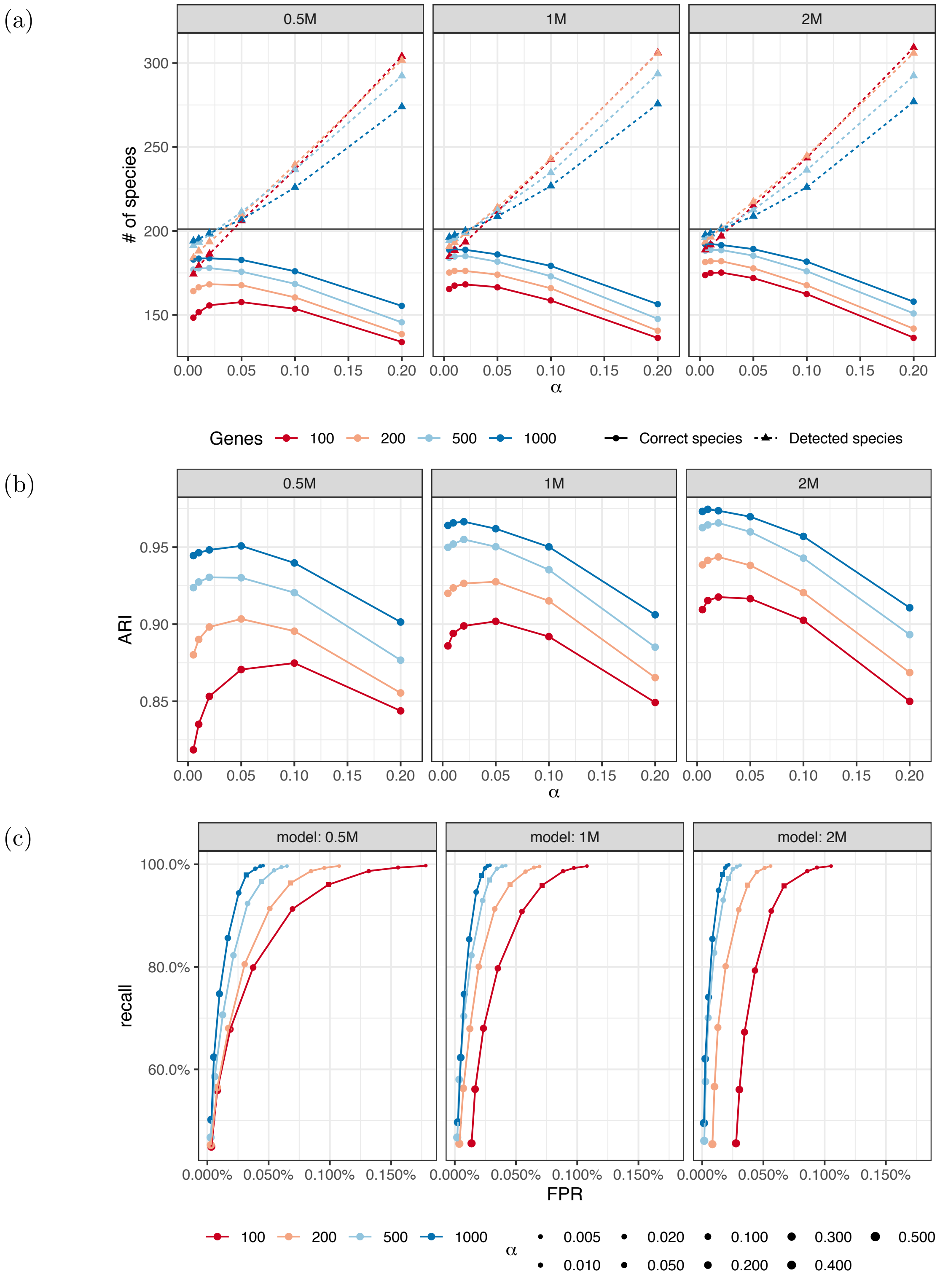

Figure 2: **Accuracy of SODA on the large dataset.** (a) We show the number of species that are completely correctly delimited (solid lines) and the total number of species found by SODA (dashed lines). Results divided into three model conditions with very high ILS (0.5M), high ILS (1M) and moderate ILS (2M) as we change α (x-axis) and the number of genes (colors). We clip α at 0.2 but show full results in Figure S3. (b) ARI (y-axis) shows the accuracy of SODA. (c) ROC showing recall versus False Positive Rate (FPR) for all model conditions and different choices of α (dot size). The default value shown as a square.

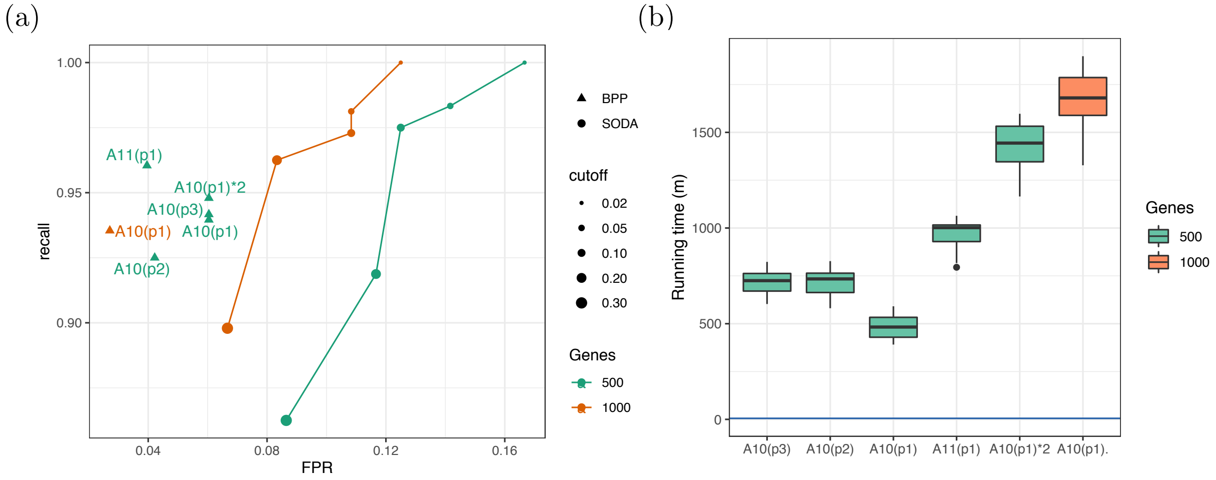

Figure 3: **Results on the small dataset.** (a) ROC curves for SODA and BPP on the small 4-taxon dataset with 500 or 1000 genes (colors) averaged over all replicates. BPP with 500 genes has several settings: three prior values for θ_s are p1=IG(1.525,0.0001), p2=IG(1.525,0.001) and p3=IG(1.525,0.01); A11 indicates species tree co-estimation while A10 indicates using ASTRAL as the guide tree; A10(p1)*2 indicates doubling the number of MCMC iterations. (b) The running time of BPP with various settings. Blue horizontal line shows the running time of SODA, including gene tree estimation. Both methods are run on Intel Xeon E5-2680v3 processors; however, SODA uses one core while we ran BPP with 4 threads and 4 cores.

11% with 1000 genes. Changing α trades off FPR and recall in expected ways; e.g., with $\alpha = 0.02$, recall is 100% but FPR increases to 12% for 1000 genes and 17% for 500 genes. Compared to SODA, BPP has a lower FPR, ranging between 4% and 6% for 500bp and 3% for 1000 genes. However, the recall of BPP is not better than default SODA and ranges between 92% and 96%, depending on the setting used. Just like SODA, increased number of genes improves FPR of BPP but not its recall. Overall, SODA-default seems to err on the side of combining individuals, while BPP tends to over-split species. Judging by the ARI (Table 1) BPP has better accuracy overall; e.g., SODA-default has an ARI of 0.78 on 500 genes while ARI of BPP ranges between 0.85 and 0.89. Overall, the parameter choices for BPP do impact accuracy but not in major ways. Doubling the number of iterations has limited to no impact and the two choices for the prior were almost identical. A third setting resulted in lower FPR but also lower recall (Fig. 3a). The only parameter that increases accuracy substantially is the switch to species tree co-estimation, which improves recall by 3.5% and reduces FPR by 0.2%.

The slightly higher accuracy of BPP comes at a steep price in running time (Fig. 3b). BPP takes between 400 and 1900 minutes on these data, given 4 cores. In contrast, SODA never takes more than a minute. The gene tree estimation also takes a few minutes (≈ 5) for this dataset. Deviating from our default setting further increases the running time of BPP, with little impact on accuracy. For example, doubling the number of iterations results in a 3× increase in running time and asking BPP to co-estimate the species tree results in a 2× increase.

Table 1: ARI on real data. We show mean (standard deviation) across replicates.

	SODA		BPP				
Genes	0.05	0.1	A10(p1)	A10(p2)	A10(p3)	A10(p1)*2	A11(p1)
500	0.75 (0.26)	0.78 (0.25)	0.85 (0.20)	0.86 (0.16)	0.85 (0.19)	0.86 (0.19)	0.89 (0.15)
1000	0.80 (0.23)	0.79 (0.23)	0.90 (0.12)	-	-	-	-

5 Discussion

We designed SODA, an ultra-fast and yet accurate method for species tree delimitation. SODA relies on frequencies of quartet topologies to decide whether each branch in a guide tree inferred from gene trees is likely to have strictly positive length, using results to infer an extended species tree, which readily defines species boundaries. SODA focuses exclusively on the MSC-based species delimitation, as applicable to Eukaryotes. It is not designed for defining viral quasispecies (Domingo *et al.*, 2012; Töpfer *et al.*, 2014) or dividing Prokaryotes into OTUs (Edgar, 2010).

Our method, like many of the existing methods, is based on several strong assumptions. Most importantly, it ignores the population structure within species and does not consider gene flow. Presence of gene flow or population structure can lead to over-splitting for other methods like BPP, as several recent studies demonstrate (Carstens *et al.*, 2013; Jackson *et al.*, 2017; Sukumaran and Knowles, 2017; Leaché *et al.*, 2019). We do not have any reason to believe that SODA, which is not yet tested under those conditions, is immune to gene flow and population structure.

Like other methods relying on input gene trees, the accuracy of SODA may depend on the input gene trees (Olave *et al.*, 2014). It is helpful that we only rely on unrooted tree topologies, and thus, errors in branch length and rooting do not affect SODA. Except for an outlier, we didn't detect a strong impact from gene tree error (Fig S6). Nevertheless, errors in gene trees may bias SODA towards over-splitting because gene tree error tends to increase observed discordance (Patel, 2013; Mirarab *et al.*, 2014b). Note that our simulations used gene trees with high levels of error (Fig. S6).

SODA also relies on a guide species tree. Luckily, given large numbers of genes, the accuracy of species trees tend to be much higher than gene trees, and Rabiee *et al.* (2019) showed that individuals of the same species often group together in ASTRAL trees. Nevertheless, if the guide tree includes substantial error, SODA may suffer. In our tests, switching to true guide trees resulting in small improvements in accuracy, showing guide trees are not a major source of error (Fig. S4).

We were able to compare SODA against one of the most widely-used alternatives, BPP. Previous simulation studies (e.g., Zhang *et al.*, 2011; Camargo *et al.*, 2012; Yang and Rannala, 2014b; Jackson *et al.*, 2017) and empirical analyses (Ruane *et al.*, 2013; Klein *et al.*, 2016; Hotaling *et al.*, 2016) have established BPP as the most accurate and preferred delimitation method. We do not expect other Bayesian methods to be substantially more accurate than BPP (Camargo *et al.*, 2012). And they are not much faster either. For example, STACEY took 7 days (Jones, 2017) on the Giarla and Esselstyn (2015) dataset with 19 individuals from 9 shrew species and 500 genes; SODA, on the same data, finished in a matter of seconds and produced identical results (Fig. S7). We were not able to use alternative methods that take input gene trees as input. In the case of SpedeSTEM, it requires rooted ultrametric gene trees, which we do not have (our gene trees are inferred). Using SpedeSTEM should be combined with a dating method, which makes the analyses less reliable; moreover, SpedeSTEM has been less accurate than BPP in previous analyses (Camargo *et al.*, 2012). Other methods (e.g., Brownie (O'Meara, 2010)) do not seem to have stable software support, preventing us from using them. Yet older methods based on individual loci (e.g., GMYC by Pons *et al.*, 2006) were not relevant to our multi-locus datasets.

The advantage of SODA over BPP, in our tests, was two-fold: much better scalability and somewhat better recall. BPP cannot handle more than tens of populations while SODA can easily handle 1000 populations (used in our large simulations). The relative strengths of the two methods suggest a natural way to combine them. We can first run SODA on the entire (large) dataset to obtain an initial delimitation. Given its near-perfect recall (especially with lower α), SODA will rarely break up real species but may combine several species together. Then, on each partition produced by SODA, we can run BPP to further refine the delimitation. This divide-and-conquer approach should be tested in the future.

Software availability. SODA code and data presented here are available on `https://github.com/maryamrabiee/SODA`.

References

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

Camargo, A., Morando, M., Avila, L. J., and Sites, J. W. (2012). Species delimitation with abc and other coalescent-based methods: A test of accuracy with simulations and an empirical example with lizards of the liolaemus darwinii complex (Squamata: Liolaemidae). *Evolution*.

Carstens, B. C., Pelletier, T. A., Reid, N. M., and Satler, J. D. (2013). How to fail at species delimitation.

Coyne, J. A. and Orr, H. A. (2004). *Speciation*. Sinauer Associates Sunderland, MA.

Degnan, J. H. and Salter, L. A. (2005). Gene tree distributions under the coalescent process. *Evolution*, **59**(1), 24–37.

Domingo, E., Sheldon, J., and Perales, C. (2012). Viral Quasispecies Evolution. *Microbiology and Molecular Biology Reviews*, **76**(2), 159–216.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**(19), 2460–2461.

Ence, D. D. and Carstens, B. C. (2011). SpedeSTEM: a rapid and accurate method for species delimitation. *Molecular Ecology Resources*, **11**(3), 473–480.

Esselstyn, J. A., Evans, B. J., Sedlock, J. L., Khan, F. A. A., and Heaney, L. R. (2012). Single-locus species delimitation: A test of the mixed yule-coalescent model, with an empirical application to Philippine round-leaf bats. *Proceedings of the Royal Society B: Biological Sciences*, **279**(1743), 3678–3686.

Fletcher, W. and Yang, Z. (2009). INDELible: A flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, **26**(8), 1879–1888.

Fujisawa, T. and Barraclough, T. G. (2013). Delimiting species using single-locus data and the generalized mixed yule coalescent approach: a revised method and evaluation on simulated data sets. *Systematic biology*, **62**(5), 707–724.

Giarla, T. C. and Esselstyn, J. A. (2015). The Challenges of Resolving a Rapid, Recent Radiation: Empirical and Simulated Phylogenomics of Philippine Shrews. *Systematic Biology*, **64**(5), 727–740.

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., and Francis, C. M. (2004). Identification of Birds through DNA Barcodes. *PLoS Biology*, **2**(10), e312.

Hotaling, S., Foley, M. E., Lawrence, N. M., Bocanegra, J., Blanco, M. B., Rasoloarison, R., Kappeler, P. M., Barrett, M. A., Yoder, A. D., and Weisrock, D. W. (2016). Species discovery and validation in a cryptic radiation of endangered primates: coalescent-based species delimitation in Madagascar's mouse lemurs. *Molecular Ecology*, **25**(9), 2029–2045.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, **2**(1), 193–218.

Hudson, R. R. and Coyne, J. A. (2002). Mathematical consequences of the genealogical species concept. *Evolution*, **56**(8), 1557–1565.

Huelsenbeck, J. P., Andolfatto, P., and Huelsenbeck, E. T. (2011). Structurama: Bayesian inference of population structure. *Evolutionary Bioinformatics*, **7**, EBO–S6761.

Jackson, N. D., Carstens, B. C., Morales, A. E., and O'Meara, B. C. (2017). Species delimitation with gene flow. *Systematic Biology*, **66**(5), 799–812.

Jones, G. (2017). Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *Journal of Mathematical Biology*, **74**(1-2), 447–467.

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability*, **19**(1982), 27–43.

Klein, E. R., Harris, R. B., Fisher, R. N., and Reeder, T. W. (2016). Biogeographical history and coalescent species delimitation of Pacific island skinks (Squamata: Scincidae: Emoia cyanura species group). *Journal of Biogeography*, **43**(10), 1917–1929.

Knowles, L. L. and Carstens, B. C. (2007). Delimiting Species without Monophyletic Gene Trees. *Systematic Biology*, **56**(6), 887–895.

Leaché, A. D., Fujita, M. K., Minin, V. N., and Bouckaert, R. R. (2014). Species Delimitation using Genome-Wide SNP Data. *Systematic Biology*, **63**(4), 534–542.

Leaché, A. D., Zhu, T., Rannala, B., and Yang, Z. (2019). The Spectre of Too Many Species. *Systematic Biology*, **68**(1), 168–181.

Maddison, W. P. (1997). Gene Trees in Species Trees. *Systematic Biology*, **46**(3), 523–536.

Mallo, D., De Oliveira Martins, L., and Posada, D. (2016). SimPhy: Phylogenomic Simulation of Gene, Locus, and Species Trees. *Systematic biology*, **65**(2), 334–44.

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

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014a). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**(17), i541–i548.

Mirarab, S., Bayzid, M. S., Boussau, B., and Warnow, T. (2014b). Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, **346**(6215), 1250463–1250463.

Olave, M., Solà, E., and Knowles, L. L. (2014). Upstream Analyses Create Problems with DNA-Based Species Delimitation. *Systematic Biology*, **63**(2), 263–271.

O'Meara, B. C. (2010). New Heuristic Methods for Joint Species Delimitation and Species Tree Inference. *Systematic Biology*, **59**(1), 59–73.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular biology and evolution*, **5**(5), 568–583.

Patel, S. (2013). Error in Phylogenetic Estimation for Bushes in the Tree of Life. *Journal of Phylogenetics & Evolutionary Biology*, **01**(02), 110.

Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., Kamoun, S., Sumlin, W. D., and Vogler, A. P. (2006). Sequence-Based Species Delimitation for the DNA Taxonomy of Undescribed Insects. *Systematic Biology*, **55**(4), 595–609.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, **5**(3).

Puillandre, N., Lambert, A., Brouillet, S., and ACHAZ, G. (2012). Abgd, automatic barcode gap discovery for primary species delimitation. *Molecular ecology*, **21**(8), 1864–1877.

Rabiee, M., Sayyari, E., and Mirarab, S. (2019). Multi-allele species reconstruction using ASTRAL. *Molecular Phylogenetics and Evolution*, **130**, 286–296.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, **66**(336), 846–850.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, **164**(4), 1645–1656.

Ruane, S., Bryson, R. W., Pyron, R. A., and Burbrink, F. T. (2013). Coalescent Species Delimitation in Milksnakes (Genus Lampropeltis) and Impacts on Phylogenetic Comparative Analyses. *Systematic Biology*, **63**(2), 231–250.

Sayyari, E. and Mirarab, S. (2016). Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Molecular Biology and Evolution*, **33**(7), 1654–1668.

Sayyari, E. and Mirarab, S. (2018). Testing for Polytomies in Phylogenetic Species Trees Using Quartet Frequencies. *Genes*, **9**(3), 132.

Solís-Lemus, C., Knowles, L. L., and Ané, C. (2015). Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution*, **69**(2), 492–507.

Sukumaran, J. and Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, **26**(12), 1569–1571.

Sukumaran, J. and Knowles, L. L. (2017). Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences*, **114**(7), 1607–1612.

Töpfer, A., Marschall, T., Bull, R. A., Luciani, F., Schönhuth, A., and Beerenwinkel, N. (2014). Viral Quasispecies Assembly via Maximal Clique Enumeration. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 309–310.

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

Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, **107**(20), 9264–9269.

Yang, Z. and Rannala, B. (2014a). Unguided species delimitation using dna sequence data from multiple loci. *Molecular Biology and Evolution*, **31**(12), 3125–3135.

Yang, Z. and Rannala, B. (2014b). Unguided Species Delimitation Using DNA Sequence Data from Multiple Loci. *Molecular Biology and Evolution*, **31**(12), 3125–3135.

Zhang, C., Zhang, D.-X., Zhu, T., and Yang, Z. (2011). Evaluation of a Bayesian Coalescent Method of Species Delimitation. *Systematic Biology*, **60**(6), 747–761.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, **19**(S6), 153.

Zhang, J., Kapli, P., Pavlidis, P., and Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, **29**(22), 2869–2876.

Zhang, L. and Cui, Y. (2010). An efficient method for DNA-based species assignment via gene tree and species tree reconciliation. In *International Workshop on Algorithms in Bioinformatics*, pages 300–311. Springer.