Bioinformatics, 36(24), 2020, 5623–5631
doi: 10.1093/bioinformatics/btaa1010
Advance Access Publication Date: 8 February 2021
Original Paper

Phylogenetics

SODA: multi-locus species delimitation using quartet frequencies

Maryam Rabiee[1] and Siavash Mirarab [2,*]

[1]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093, USA and [2]Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA

*To whom correspondence should be addressed.

Associate Editor: Yann Ponty

Received on March 18, 2020; revised on October 19, 2020; editorial decision on November 13, 2020; accepted on November 21, 2020

Abstract

Motivation: 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 computational evolutionary biology. While many methods exist for species delimitation, most based on the coalescent theory, few are scalable to very large datasets, and methods that scale tend to be not accurate. Species delimitation is closely related to species tree inference from discordant gene trees, a problem that has enjoyed rapid advances in recent years.

Results: In this article, we build on the accuracy and scalability of recent quartet-based methods for species tree estimation and propose a new method called SODA for species delimitation. SODA relies heavily on a recently developed method for testing zero branch length in species trees. In extensive simulations, we show that SODA can easily scale to very large datasets while maintaining high accuracy.

Availability and implementation: The code and data presented here are available on https://github.com/maryamrabiee/SODA.

Contact: smirarab@ucsd.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

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; Rannala and Yang, 2020). 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 offsprings.

A wide range of species delimitation methods exist. 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 (Esselstyn et al., 2012; Fujisawa and Barraclough, 2013; Zhang et al., 2013) 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; O'Meara, 2010; Yang and Rannala, 2010), not to mention more complex processes such as gene flow (e.g. Leaché et al., 2019). Modeling coalescence 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 happen deeper than the first opportunity, producing 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.

Existing methods for species delimitation under the MSC model tend to suffer from one of two limitations. The most accurate methods are based on Bayesian MCMC and infer gene trees, (optionally) species trees and species boundaries [e.g. BPP (Yang and Rannala, 2010, 2014), 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). For example, Musher and Cracraft (2018) had to divide their

dataset of 62 individuals into six subsets, and Oliveira et al. (2015) used a subsample of 20 out of 137 to use BPP to avoid mixing problems that usually happen with large datasets; running BPP on a dataset of 40 populations in our study needed 36 h of running time. The 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), then, 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 makes 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 do.

In this article, 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). A statistically consistent method, ASTRAL infers 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 article 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 can define species. By relying on quartet frequencies and the ASTRAL machinery, SODA is able to handle very large datasets with short running times. We first describe SODA in detail and then evaluate its accuracy and scalability in simulation and on empirical datasets.

2 Materials and methods

2.1 Coalescent-based topology-based delimitation

We take a two-step approach to species delimitation and assume unrooted gene tree *topologies* are already inferred from sequence data. Thus, we are given a set of unrooted gene trees $\mathcal{G}$ on individuals $\ell = \{l_1 \dots l_m\}$. Ignoring errors in estimated gene trees, we assume these gene trees follow the MSC model. Optionally, we are given a partition of ℓ into populations $\mathcal{P} = \{P_1 \dots P_p\}$; when $\mathcal{P}$ is not given, we define each individual as a singleton population. A partition of $\mathcal{P}$ into $\mathcal{S} = \{S_1 \dots S_n\}$ produces a mapping $r : \mathcal{P} \to \{1 \dots n\}$ and by extension $q : \ell \to \{1 \dots n\}$ where $r(x)$ and $q(y)$ give a species index for $x \in \mathcal{P}$ and $y \in \ell$.

We say a partition is coalescent-consistent if for any set $A \subset \ell$ with at most one individual from each population and all individuals mapped to the same species ($\forall_{i,j} l_i, l_j \in A : \nexists_x l_i, l_j \in P_x, \exists_y l_i, l_j \in S_y$), the distribution of $\mathcal{G}$ restricted to S is consistent with the neutral Kingman (1982) coalescence process. Because Kingman's process is robust to subsampling, if a partition is coalescent-consistent, further breaking each species into smaller species would remain coalescent-consistent. Having two species that could be combined without violating the coalescent model is not justified under the MSC because the model provides no support for the division. Thus, we formulate coalescent-based species delimitation as the problem of finding a coalescent-consistent partition that is not a refinement of any other coalescent-consistent partitions. Our method seeks to solve this problem within further restrictions stated below.

This formulation follows the MSC model (free coalescent of lineages within species but constrained coalescence across the species) and shares its assumptions. The only population structure within the species that is modeled is the given structure $\mathcal{P}$, which is known *apriori*. Thus, it assumes that individuals selected from different populations of the same species evolve according to the neutral Wright-Fisher model, resulting in a distribution of gene trees within a species that follows the Kingman (1982) coalescent process. This assumption is most defensible when the populations within a species do not further have *strong* differentiation. MSC also assumes lineages sampled from different species do not coalesce more recently than their separation event (Supplementary Fig. S8); thus, it ignores gene flow across species. While these assumptions can all be violated on real data, they provide a useful model that allows fast delimitation. We revisit these assumptions in the discussion session.

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). Simultaneously dealing with these two processes is challenging, 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. To do so, we rely on the distribution of gene tree topologies under the MSC model, in particular for quartets of species (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 a single individual, and a single 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

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 ℓ, built by adding to each leaf of T^* all individuals corresponding to that

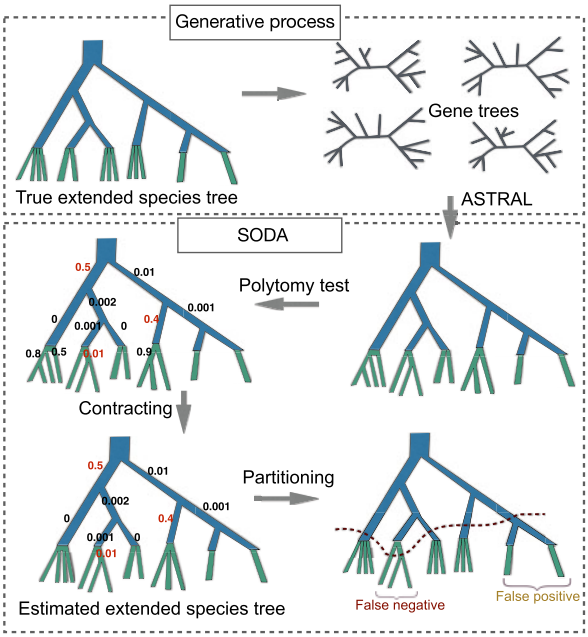

Fig. 1. True extended species tree generates gene trees under the MSC model. From unrooted gene tree topologies (inferred from sequence data), SODA first estimates a guide tree using ASTRAL and then test the null hypothesis that each branch has length zero, obtaining a *P*-value (middle left), which may result in FP or FN rejection or retention of the null (red *P*-value s). SODA then contracts branches where the null is retained as long as contracting them does not contradict with the monophyly of species defined by branches where the null hypothesis is rejected; thus, we keep some branches with high *P*-value (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 hypothesis testing (e.g. *P*-value s 0.5 and 0.4) do not result in erroneous delimitation.

Algorithm 1 - SODA Algorithm

function SODA($\mathcal{G}$, T,α)
if guide tree T is not given **then**
$T \leftarrow$ run ASTRAL on $\mathcal{G}$
end if
for internal branch e of T **do**
$p(e) \leftarrow$ P-value of the null hypothesis that length(e) = 0
end for
Root T on the $\arg\min_e p(e)$
ContractToExtendedSpeciesTre(T,α)
partition $\mathcal{P}$ by cutting all 'keep' internal edges of T
end function
function ContractToExtendedSpeciesTree(T,α)
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'
end if
end for
for $e \in$ internal edges of T not marked 'keep' **do**
Contract e
end for
end function

species as a polytomy (Fig. 1); i.e. for leaf $s \in \mathcal{S}$ of T, add a child for every $r \in q^{-1}(s) \subset \ell$. When populations are known *apriori*, we can similarly define the extended species tree $\mathcal{T}$ as a rooted tree labeled by $\mathcal{P}$, built by adding to each leaf of T^* all populations corresponding to that species as a polytomy.

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 significance level α, described below. SODA first infers (or takes as input) a guide tree T. The guide tree, a concept introduced by Yang and Rannala (2010), is a phylogenetic tree with leaves set to known populations (i.e the most divided possible delimitation). SODA assumes T is a *resolution* of the true extended species tree $\mathcal{T}$, meaning that it includes all branches of the species tree and has arbitrary relationships between populations of the same species. SODA then contracts *some of the* branches of T as long as it cannot reject the null hypothesis that they have length zero under the MSC model using α as the confidence level for statistical tests; the contracted tree is an estimated extended species tree $\hat{\mathcal{T}}$. Finally, SODA cuts internal branches of the $\hat{\mathcal{T}}$ to cluster species (Fig. 1). Below, we describe each step in more detail.**Guide Tree:** SODA needs a (potentially unrooted) guide tree T on the leafset $\mathcal{P}$. If not provided by the user, we infer the guide tree by running ASTRAL-III on $\mathcal{G}$; note that using the multi-individual version of ASTRAL (Rabiee *et al.*, 2019), we can ensure that the tree generated by ASTRAL is labeled by $\mathcal{P}$, as opposed to the original individuals (if different). 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 next test the null hypothesis that it has zero length in 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 (Allman *et al.*, 2011; Pamilo and Nei, 1988). Moreover, gene trees are assumed independent. Thus, under the null hypothesis of a polytomy, the frequency of quartet topologies around each branch should follow a multinomial distribution with three categories, each with probability $\frac{1}{3}$. Whether observed frequencies of the three possible topologies for a quartet follow an equiprobable multinomial distribution can be tested using a Chi-Squared test, which is what the method of Sayyari and Mirarab (2018) uses. To achieve scalability, this method treats branches independently [based on the locality assumption of Sayyari and Mirarab (2016)] and takes the average of quartet frequencies for all quartets around each branch (which can be done easily in $O(n^2k)$ time). The test assumes the input gene tree set is an error-free random sample generated by the MSC model from the true species tree. It 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 at a user-specified level α. 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). Parameter α can be adjusted for controlling how aggressively SODA divides species. Increasing α results in rejecting more null hypotheses and hence, dividing individuals into more species. To ensure that we can get a valid extended species tree (with monophyletic species), we need polytomies to form only above the 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 $\hat{\mathcal{T}}$, the partition is obtained by cutting the remaining internal branches of $\hat{\mathcal{T}}$. The partition produced would be identical if we only cut internal branches that have at least one terminal branch as a child.

The accuracy of the algorithm, in addition to assumptions made by our problem formulation, depends on the accuracy of the statistical test. We formalize this notion in three claims (proofs in Supplementary Material).

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 ($\hat{\mathcal{T}}=\mathcal{T}$). Additionally, it will correctly delimitate species if all species are sampled more than once.*

This claim provides a reassuring 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 tree $\mathcal{T}$, 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 the division of a species; however, an FP does not always divide S into exactly two parts. For example, if T has a caterpillar (a.k.a ladder-like) topology on S, an FP on the branch above the cherry (i.e. a node with two leaf children) 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. 1) S_1 and S_2 each have one sampled individuals and form a cherry. 2) 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 Experimental 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 conditions where other methods cannot run, whereas the other dataset is small and enables us to compare SODA 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 five 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 (i.e. the proportion of quartet trees that are shared between the two trees) is 0.78, 0.62 and 0.50 for these model conditions (Supplementary Fig. S5). Note that a quartet score of $\frac{1}{3}$ indicates random trees, and even a value of 0.5 indicates very high levels of ILS. 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 gene trees are estimated using FastTree (Price et al., 2010) from the alignments; estimated gene trees are used throughout the experiments in this article. The average gene tree error, measured as normalized Robinson and Foulds (1981) (RF) distance for the three model conditions is 0.25, 0.31 and 0.42 with large variance (Supplementary Fig. S5).

Small dataset: We simulate a new dataset using SimPhy with 20 replicates, each only four species, ten individuals per species and 1000 genes per replicate (commands shown in Appendix 1). 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 the birth-only model with rate = 0.00001. These settings lead to a high level of ILS, capturing a scenario where species delimitation is challenging. 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 (Supplementary Table S1). We simulate 1000 bp alignments on each gene tree using INDELible (Fletcher and Yang, 2009) and estimate gene trees using FastTree based on the alignments (Price *et al.*, 2010). The average gene tree error (normalized RF between true and estimated gene trees) is 43%. Gene tree distance between pairs of individuals of the same species has a wide range in our simulations, ranging between 10^{-5} and 10^{-2} mutations per site in most cases (Supplementary Fig. S2).

Empirical dataset: We study three biological datasets. To show applicability on large data, we use the dataset of Protea L. with recent radiations (Mitchell et al., 2017), which has sampled multiple individuals from 59 species of Protea and six outgroup species (a total of 163 tips) and obtained 498 low-copy, orthologous nuclear loci. Due to its size, this dataset is only analyzed using SODA. To be able to compare to other methods, we use a smaller dataset of lizards of the Australian wet tropics (AWT). Singhal et al. (2018) analyzed genetic data for individuals from three species groups (*Carlia rubrigularis*, *Lampropholis coggeri* and *Lampropholis robertsi*) that split into 13 putative lineages. In total, there are 25 individuals, and 3320 loci across all individuals are sequenced using an exome capture approach. Gene trees and the species tree are estimated using STARBEAST2 v0.13.5 (Ogilvie et al., 2017), and delimitation results using STACEY (Jones, 2017) and BPP are available from the original study. We also study the human dataset analyzed by Jackson et al. (2017) comprising sequences from 50 loci (415–960 bp long) for four widely sampled groups of humans defined geographically (as the original study states: these four groups are ethnically diverse and are not 'populations' in any biological sense). The dataset includes ten samples from each of Africa, Europe and Asia, and 12 samples from South and Central America. Analyzing the human dataset, where we clearly know all individuals belong to one species, enables us to test the propensity of the method to lead to false positive delimitation.

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 a Receiver Operating Characteristic (ROC) curve as we change the α setting 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 the expected similarity of pairs according to a generalized hypergeometric distribution that controls for the number of objects and classes (Hubert and Arabie, 1985). ARI equals one only for the correct partition and is close to zero for a random partition.

3.3 Methods compared

BPP. We compare SODA to the widely used Bayesian Phylogenetics and Phylogeography (BPP) method. 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 take advantage of its multi-thread version to be able to run it with up to 1000 genes. Provided with all 40 individuals of the small dataset that we sampled as 40 separate populations, BPP could not run to completion in 36 h, perhaps because the set of possible delimitations was too large. To be able to test BPP, we use two subsampled sets. First, we randomly sample 4 individuals per species and designate each as its own population, for a total of 16 populations. In the second scenario, from each species, we sample 7 individuals and randomly assigned them to 3 populations of sizes 2, 2 and, 3, comprising 12 populations in total. We use a uniform prior across all possible partitions on the resulting sets of 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 the guide tree 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. The mean of the distribution is set based on the true average of species pairwise distances in the gene trees. An example of a control file used for running BPP is given in Supplementary Figure S1 for the full set of parameters.

SpedeSTEM. Given a set of rooted ultrametric gene trees, SpedeSTEM uses the STEM (Kubatko et al., 2009) algorithm to calculate the maximum likelihood species tree considering possible species tree and delimitation combinations and uses AIC to select among the models. SpedeSTEMv0.9 software includes a pipeline for inferring ultrametric rooted gene trees using paup* (Swofford, 2001). We used SpedeSTEMv0.9 to infer the rooted ultrametric input gene trees, which we then fed to the SpedeSTEMv2 software as input. We ran delimitation using theta values of 0.1 and 0.01, and a sampling ratio of 1 (no subsampling). SpedeSTEMv2 also requires a putative assignment of populations to species. For this, we tested assigning all populations to the same putative species or assigning all 12 populations to individual species. Results were similar, and we report the latter strategy.

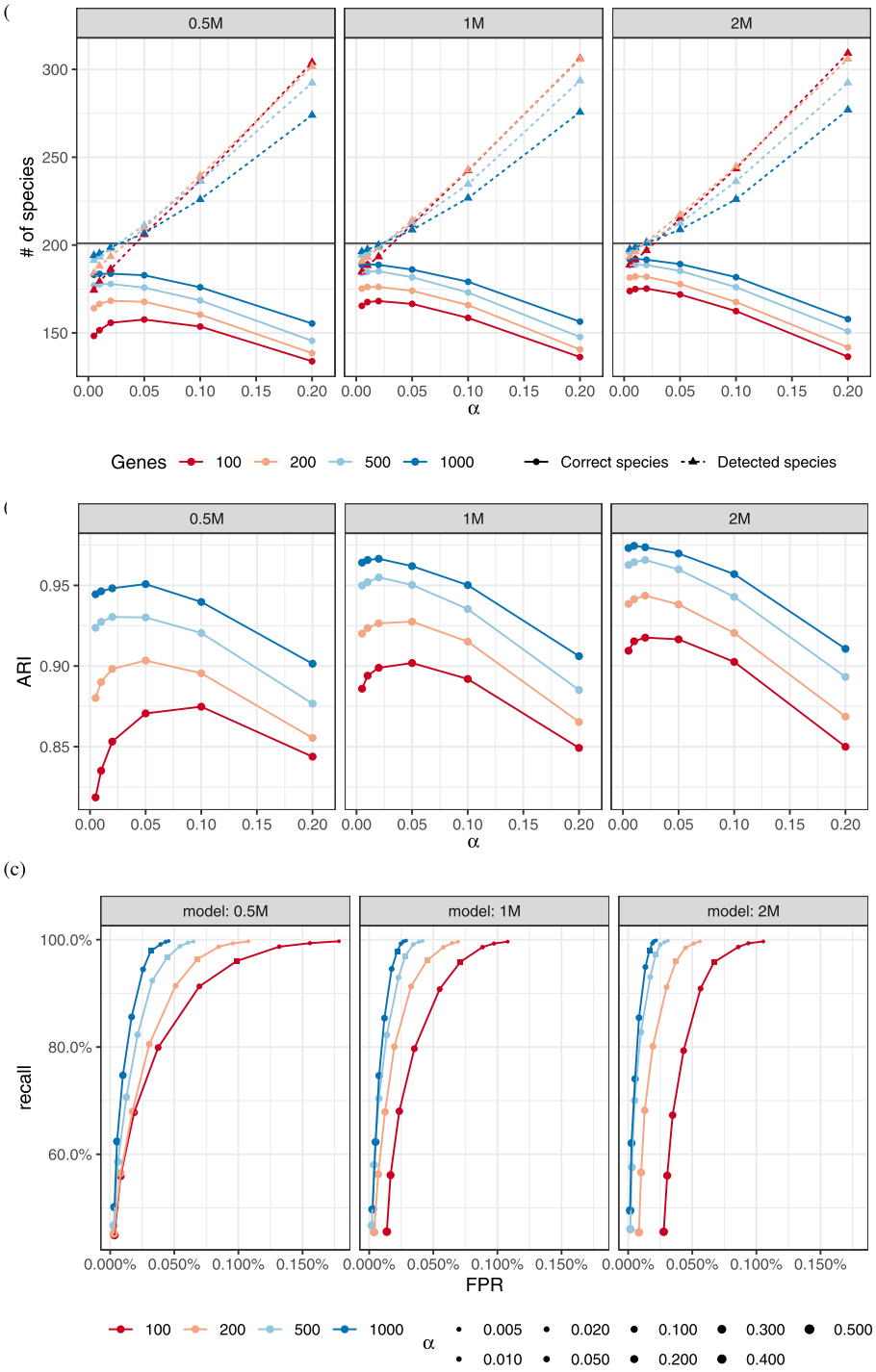

Fig. 2. Accuracy of SODA on the large dataset. (**a**) We show the number of species that are completely correctly delimited (solid lines) and the total number of species found by SODA (dashed lines). Results divided into three model conditions with very high ILS (0.5M), high ILS (1M) and moderate ILS (2M) as we change α (x-axis) and the number of genes (colors). We clip α at 0.2 but show full results in Supplementary 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

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 study 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. For tests with known populations, we inferred the guide tree using ASTRAL-multi, mapping leaves of the same population to a 'species,' assigning each population to a separate species.

4 Results

4.1 Large simulated dataset

On the large dataset with 1001 individuals, SODA takes no more than 35 min (Supplementary 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).

Some of the mistakes made by SODA are related to the guide tree. The ASTRAL tree failed to recover on average 5, 3 and 3 species as a monophyletic clade in these three model conditions, and these species could never be recovered correctly by SODA. On average, the estimated guide tree by ASTRAL (on 1000 genes) missed 4%, 3% and 2% of branches in the true extended species tree for high, moderate and low levels of ILS. The *estimated* extended species tree that SODA outputs has an RF distance of 6%, 5% and 4% to the true extended species tree.

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. 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. 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 (Supplementary Fig. S6). However, the guide tree and the extended species tree *are* impacted by increased gene tree error (Supplementary Fig. S7). As the increased error does not impact delimitation, the increased error of the guide tree must be concentrated on deep branches that do not impact delimitation.

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. Increasing α reduces FPR; however, for $\alpha > 0.1$, we observe only small gains in FPR but precipitous declines in the recall. Thus, as observed earlier, $\alpha > 0.1$ does not seem advisable. Beyond the default value, a choice of $\alpha = 0.01$ seems 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 (Supplementary Fig. S4).

4.2 Small simulated dataset

On the small dataset with four species and 16 populations, SODA (default) has 98% recall with both 500 and 1000 genes (Fig. 3a). Increasing the number of genes mostly reduces FPR, from 14% with 500 genes to 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 500 genes and 3% for 1000 genes. However, the recall of BPP is not better than the default SODA and ranges between 92% and 96%, depending on the setting used. Just like SODA, an increased number of genes improves the 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

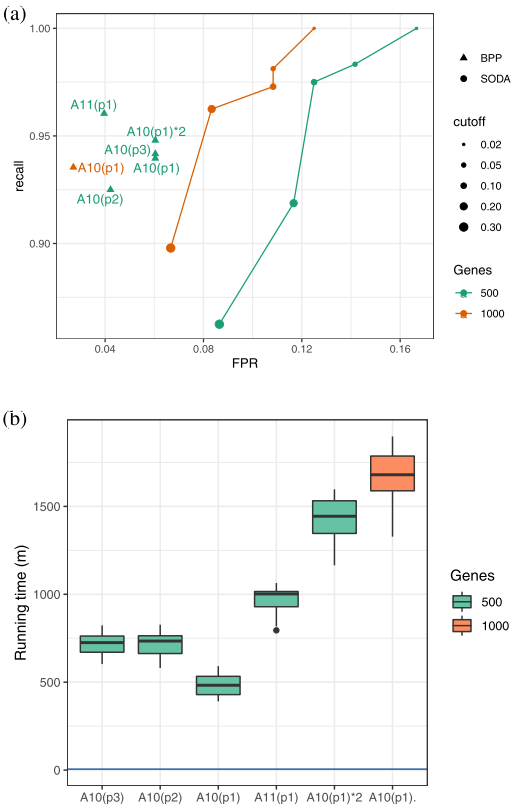

Fig. 3. Results on the small dataset. (**a**) ROC curves for SODA and BPP on the small 4-taxon dataset with 500 or 1000 genes (colors) averaged over all replicates. BPP with 500 genes has several settings: three prior values for θ_s are p1 = IG(1.525,0.0001), p2 = IG(1.525,0.001) and p3 = IG(1.525,0.01). 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. The 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

almost identical. A third setting resulted in lower FPR but also lower recall (Fig. 3a). The only parameter that increases accuracy substantially is 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 min on these data, given four cores. In contrast, SODA never takes more than a minute, and the gene tree estimation takes a few minutes ($\approx$5) for this dataset. Deviating from our default setting further increases the running time of BPP, with little impact on accuracy. For example, doubling the number of iterations results in a 3$\times$ increase in running time, and asking BPP to co-estimate the species tree results in a 2$\times$ increase.

We next compare SODA to BPP and SpedeSTEM on the setting where for each species, seven individuals divided into three populations known *apriori* are given (a total of 12 populations). In this setting, the FP rate of SODA is 8% on average, showing that species are sometimes combined together; in contrast, FN is 0, meaning that over-splitting does not occur (Table 2). BPP outperforms SODA in terms of FP (1–2%) but also occasionally over-splits (FN > 0). Overall, according to ARI, BPP remains somewhat more accurate. SpedeSTEM, in comparison, has much lower accuracy; in almost all cases, it detects exactly two species (instead of four), leading to very high FP rates and low ARI. We note that in the STEM species tree inferred, the species are often non-monophyletic.

Table 1. ARI on small datasets

	SODA			BPP				
Genes	0.01	0.02	0.05	A10(p1)	A10(p2)	A10(p3)	A10(p1)*2	A11(p1)
500	0.70 (0.24)	0.72 (0.25)	0.75 (0.26)	0.85 (0.20)	0.86 (0.16)	0.85 (0.19)	0.86 (0.19)	0.89 (0.15)
1000	0.78 (0.26)	0.79 (0.26)	0.80 (0.23)	0.90 (0.12)	–	–	–	–

Note: We show mean (standard deviation) across replicates. ARI of SODA has been measured with two thresholds (0.05 and 0.1).

Table 2. ARI on small datasets with individuals assigned to populations

	SODA	BPP		SpedeSTEM	
	0.05	A10	A11	$\theta = 0.1$	$\theta = 0.01$
FP	0.08 (0.12)	0.01 (0.02)	0.02 (0.03)	0.65 (0.11)	0.78 (0.11)
FN	0	0.01 (0.03)	0.01 (0.02)	0.03 (0.01)	0.0 (0.0)
ARI	0.82 (0.22)	0.93 (0.12)	0.89 (0.13)	0.03 (0.11)	0.05 (0.12)

Note: Three populations per species were defined randomly with 3, 2 and 2 individuals per population. We show mean (standard deviation) across replicates. A11 indicates species tree co-estimation while A10 indicates using ASTRAL as the guide tree.

Table 3. Delimitation accuracy measured using ARI for the Protea dataset with cutoff thresholds 0.01, 0.02 and 0.05

Species	Gene trees	0.01	0.02	0.05
All species	Fully resolved	0.662	0.655	0.611
	≤5 BS contracted	0.653	0.654	0.637
	≤10 BS contracted	0.654	0.654	0.628
Monophyletic species	Fully resolved	0.776	0.767	0.717
	≤5 BS contracted	0.736	0.775	0.765
	≤10 BS contracted	0.736	0.785	0.755

Note: 'Monophyletic species' means individuals of the species that make them non-monophyletic in the ASTRAL tree are pruned (19 in total). We run SODA on fully resolved gene trees and gene trees with branches with BS support below 5% or 10% collapsed.

4.3 Empirical dataset

The ARI of SODA on the Protea dataset ranges from 0.61 to 0.66 with different values for α, given all 498 gene trees available (Table 3). This dataset includes many singleton species (12 out of 59 ingroups and 5 out of 6 outgroup species) along with many non-monophyletic clades; to test the accuracy of delimitation (as opposed to species tree inference), we removed individuals that form non-monophyletic clades (19 in total) and gained better delimitation with ARI ranging from 0.72 to 0.78 on the pruned species tree.

SODA-default detects 89 species and is thus over-splitting some species. Some of the split species seem to capture populations within species. For example, *Protea Acaulos* is divided into two groups that coincide with grouping based on the geographical locations of the samples (provided by the original study); these geographical separations may reflect two sub-population of this species.

We also tested collapsing low bootstrap support branches in gene trees to deal with the effects of high gene tree estimation error. Collapsing very low support branches improved the results slightly with the default $\alpha = 0.05$ (Table 3). The improvement is more clear when we prune non-monophyletic species.

Human. SODA results support grouping all individuals into one species, and all the *P*-values across the tree are above 0.34. Thus, on this dataset, reassuringly, SODA avoids a false positive breakup to multiple species. While the recovery of humans (a relatively recent species) as one species may seem an easy case of species delimitation, Jackson *et al.* (2017) showed that BPP supports a four-species model with high posterior in all their ten replicate runs, regardless of the prior used. The authors attributed this error to population structure and used this as a motivation to introduce the PHRAPL method, which unlike BPP, did unambiguously recover humans as a single species.

Lizards. Statistical species delimitation using both BPP and STACEY support a speciation event at every node of the guide tree (regardless of priors chosen). SODA, similar to BPP and STACEY, detects all lineages as separate species with *P*-value≈0 for all branches except one branch, which does not result in a false positive (Supplementary Fig. S9). Singhal *et al.* (2018) report that the delimitation into 13 groups does not match a clear morphological separation between species, making this a potential case of a cryptic species. However, it should be noted that in the light of the results from humans, a false positive delimitation by all methods cannot be ruled out.

5 Discussion

We designed SODA, an ultra-fast and relatively 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 then 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).

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. The presence of gene flow across species 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; Leaché *et al.*, 2019; Sukumaran and Knowles, 2017). Our results on the Protea dataset indicate that SODA can suffer from a similar blindspot. We did not directly test SODA under simulation conditions with gene flow and population structure; the fact that we simulate under MSC and test under MSC can describe why our errors tend to be of over-splitting nature on simulated data. On the real Protea dataset, species were split (often by geography), showing that SODA can be sensitive to population structure within species. SODA, unlike BPP, avoided breaking humans into multiple species; however, it is hard to know whether this result is due to the presence of only 50 gene trees in this dataset or some level of robustness to population structure. Note that SODA shares its sensitivity to structure and gene flow with methods purely based on MSC. Thus, we suggest that for datasets where high levels of gene flow after speciation is probable, the results of SODA should be used as a guide to enable

more time-consuming delimitation using methods that do consider gene flow and are more robust to structure.

In addition to gene flow and population structure, several factors need to be kept in mind when using SODA. 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. Plotting accuracy of SODA versus mean gene tree error across simulation conditions, we did not detect a strong impact from gene tree error except for an outlier (Supplementary Fig. S6). Nevertheless, errors in gene tree topologies may bias SODA toward over-splitting because gene tree error tends to increase observed discordance (Mirarab et al., 2014b; Patel, 2013). Moreover, the polytomy test used by SODA makes several assumptions, including the independent treatment of branches. These assumptions could, in theory, further impact the method in the presence of many adjacent short branches. To make sure this is not the case, our simulations used gene trees with high levels of error (Supplementary Fig. S6) and included conditions with many adjacent short branches, and yet showed positive results. Nevertheless, simulations are by nature limited, and thus, further work may reveal other conditions where the method fails.

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. In our simulations, switching to true guide trees resulted in small improvements in accuracy (Supplementary Fig. S4). Nevertheless, if the guide tree includes substantial levels of error, SODA may suffer. For example, on the Protea dataset, several individuals were placed far from their presumed species. Assuming these individuals were correctly identified, we have to conclude the ASTRAL tree had several errors, a problem that SODA is not able to overcome. Finally, SODA requires that the analysis includes at least two individuals from each species, another factor that may limit its application to practice. Due to these caveats, applying SODA has to be done with care.

We were able to compare SODA against one of the most widely used alternatives, BPP. Previous simulation studies (e.g. Camargo *et al.*, 2012; Jackson *et al.*, 2017; Yang and Rannala, 2014; Zhang et al., 2011) and empirical analyses (Hotaling et al., 2016; Klein et al., 2016; Ruane et al., 2014) have established BPP as the most accurate and preferred MSC-based 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 seven 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 (Supplementary Fig. S10). In the case of SpedeSTEM, it requires rooted ultrametric gene trees, which cannot be inferred using the standard models of sequence evolution. Using SpedeSTEM should be combined with rooting and a rate model, which can make the analyses sensitive to errors in those steps; moreover, SpedeSTEM has been less accurate than BPP in previous analyses (Camargo *et al.*, 2012). In our analyses, SpedeSTEM was the least accurate method. Perhaps the lack of accuracy is due to divergences from a strict molecular clock used in our Simphy simulations, which perhaps the SpedeSTEMv0.9 default pipeline could not overcome. STEM, used in SpedeSTEM, has been shown to have low accuracy in computing the species tree (Leaché and Rannala, 2011), especially given variations in mutational processes (Huang et al., 2010).

We were not able to use other methods that take gene trees as input. For example, the algorithm of O'Meara (2010) infers species delimitation either using the full gene tree likelihood calculation, which is slow (Degnan and Salter, 2005) or using the MDC cost (Maddison, 1997). However, this method (Brownie) does not seem to currently have stable software support. Similarly, the method of Zhang and Cui (2010) relies on a species tree and partially labeled gene trees of individuals of several species; however, this method does not have a publicly available implementation. Older methods based on individual loci (e.g. GMYC by Pons *et al.*, 2006) were not relevant to our multi-locus datasets. Our attempts to run PHRAPL, which focuses on gene flow, were unsuccessful because the method did not finish after 48 h of running time.

The advantage of SODA over BPP, in our tests, was two-fold: much better scalability and slightly better recall. BPP cannot handle more than tens of populations, while SODA can easily handle 1000 populations (used in our large simulations). However, overall, BPP was more accurate, especially when allowed to co-estimate the species tree. 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. The results of SODA can be used to define populations and to divide the dataset into smaller subsets for a more extensive BPP analysis. This divide-and-conquer approach is what many analyses use in practice (e.g. Musher and Cracraft, 2018) using a manual curation; SODA can help automate that process.

Financial Support: This work was supported by the NSF [IIS-1845967 to S.M. and M.R.]. Computations were performed on the San Diego Supercomputer Center (SDSC) through XSEDE allocations, which is supported by the NSF [ACI-1053575].

Conflict of Interest: none declared.

References

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

Camargo,A. *et al.* (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*, **66**, 2834–2849.

Carstens,B.C. *et al.* (2013) How to fail at species delimitation. *Mol. Ecol.*, **22**, 4369–4383.

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**, 24–37.

Domingo,E. *et al.* (2012) Viral quasispecies evolution. *Microbiol. Mol. Biol. Rev.*, **76**, 159–216.

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

Esselstyn,J.A. *et al.* (2012) Single-locus species delimitation: a test of the mixed Yule-coalescent model, with an empirical application to Philippine round-leaf bats. *Proc. R. Soc. B Biol. Sci.*, **279**, 3678–3686.

Fletcher,W. and Yang,Z. (2009) INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.*, **26**, 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. *Syst. Biol.*, **62**, 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. *Syst. Biol.*, **64**, 727–740.

Hebert,P.D.N. *et al.* (2004) Identification of birds through DNA barcodes. *PLoS Biol.*, **2**, e312.

Hotaling,S. *et al.* (2016) Species discovery and validation in a cryptic radiation of endangered primates: coalescent-based species delimitation in Madagascar's mouse lemurs. *Mol. Ecol.*, **25**, 2029–2045.

Huang,H. *et al.* (2010) 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.*, **59**, 573–583.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

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

Huelsenbeck,J.P. *et al.* (2011) Structurama: Bayesian inference of population structure. *Evol. Bioinf.*, **7**, EBO.S6761.

Jackson,N.D. *et al.* (2017) Species delimitation with gene flow. *Syst. Biol.*, **66**, 799–812.

Jones,G. (2017) Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *J. Math. Biol.*, **74**, 447–467.

Kingman,J.F.C. (1982) On the genealogy of large populations. *J. Appl. Probab.*, **19**, 27–43.
Klein,E.R. *et al.* (2016) Biogeographical history and coalescent species delimitation of Pacific island skinks (Squamata: Scincidae: *Emoia cyanura* species group). *J. Biogeography*, **43**, 1917–1929.
Knowles,L.L. and Carstens,B.C. (2007) Delimiting species without monophyletic gene trees. *Syst. Biol.*, **56**, 887–895.
Kubatko,L.S. *et al.* (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, **25**, 971–973.
Leaché,A.D. and Rannala,B. (2011) The accuracy of species tree estimation under simulation: a comparison of methods. *Syst. Biol.*, **60**, 126–137.
Leaché,A.D. *et al.* (2014) Species delimitation using genome-wide SNP data. *Syst. Biol.*, **63**, 534–542.
Leaché,A.D. *et al.* (2019) The Spectre of too many species. *Syst. Biol.*, **68**, 168–181.
Maddison,W.P. (1997) Gene trees in species trees. *Syst. Biol.*, **46**, 523–536.
Mallo,D. *et al.* (2016) SimPhy: phylogenomic simulation of gene, locus, and species trees. *Syst. Biol.*, **65**, 334–344.
Mirarab,S. and Warnow,T. (2015) ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, **31**, i44–i52.
Mirarab,S. *et al.* (2014a) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**, i541–i548.
Mirarab,S. *et al.* (2014b) Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, **346**, 1250463–1250463.
Mitchell,N. *et al.* (2017) Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of protea l. *Am.J. Bot.*, **104**, 102–115.
Musher,L.J. and Cracraft,J. (2018) Phylogenomics and species delimitation of a complex radiation of Neotropical suboscine birds (Pachyramphus). *Mol. Phylogenet. Evol.*, **118**, 204–221.
Ogilvie,H.A. *et al.* (2017) Starbeast2 brings faster species tree inference and accurate estimates of substitution rates. *Mol. Biol. Evol.*, **34**, 2101–2114.
Olave,M. *et al.* (2014) Upstream analyses create problems with DNA-based species delimitation. *Syst. Biol.*, **63**, 263–271.
Oliveira,E.F. *et al.* (2015) Speciation with gene flow in whiptail lizards from a neotropical xeric biome. *Mol. Ecol.*, **24**, 5957–5975.
O'Meara,B.C. (2010) New heuristic methods for joint species delimitation and species tree inference. *Syst. Biol.*, **59**, 59–73.
Pamilo,P. and Nei,M. (1988) Relationships between gene trees and species trees. *Mol. Biol. Evol.*, **5**, 568–583.
Patel,S. (2013) Error in phylogenetic estimation for bushes in the tree of life. *J. Phylogenet. Evol. Biol.*, **01**, 110.
Pons,J. *et al.* (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Syst. Biol.*, **55**, 595–609.
Price,M.N. *et al.* (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
Puillandre,N. *et al.* (2012) Abgd, automatic barcode gap discovery for primary species delimitation. *Mol. Ecol.*, **21**, 1864–1877.
Rabiee,M. *et al.* (2019) Multi-allele species reconstruction using ASTRAL. *Mol. Phylogenet. Evol.*, **130**, 286–296.
Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, **66**, 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**, 1645–1656.
Rannala,B. and Yang,Z. (2020) Species delimitation.
Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.
Ruane,S. *et al.* (2014) Coalescent species delimitation in milksnakes (Genus *Lampropeltis*) and impacts on phylogenetic comparative analyses. *Syst. Biol.*, **63**, 231–250.
Sayyari,E. and Mirarab,S. (2016) Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.*, **33**, 1654–1668.
Sayyari,E. and Mirarab,S. (2018) Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes*, **9**, 132.
Singhal,S. *et al.* (2018) A framework for resolving cryptic species: a case study from the lizards of the Australian wet tropics. *Syst. Biol.*, **67**, 1061–1075.
Solís-Lemus,C. *et al.* (2015) Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution*, **69**, 492–507.
Sukumaran,J. and Holder,M.T. (2010) DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, **26**, 1569–1571.
Sukumaran,J. and Knowles,L.L. (2017) Multispecies coalescent delimits structure, not species. *Proc. Natl. Acad. Sci. USA*, **114**, 1607–1612.
*Swofford,D.L. (2001) Paup: Phylogenetic analysis using parsimony (and other methods) 4.0. b5.
Töpfer,A. *et al.* (2014) Viral quasispecies assembly via maximal clique enumeration. *PLoS Comput Biol*, **10**, e1003515.
Xu,B. and Yang,Z. (2016) Challenges in species tree estimation under the multispecies coalescent model. *Genetics*, **204**, 1353–1368.
Yang,Z. and Rannala,B. (2010) Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. USA*, **107**, 9264–9269.
Yang,Z. and Rannala,B. (2014) Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.*, **31**, 3125–3135.
Zhang,C. *et al.* (2011) Evaluation of a Bayesian coalescent method of species delimitation. *Syst. Biol.*, **60**, 747–761.
Zhang,C. *et al.* (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, **19**, 153.
Zhang,J. *et al.* (2013) A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, **29**, 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*. Springer, Berlin, Heidelberg, pp. 300–311.