Syst. Biol. 74(4):656–671, 2025

https://doi.org/10.1093/sysbio/syaf017
Advance Access Publication May 30, 2025

PickMe: Sample Selection for Species Tree Reconstruction using Coalescent Weighted Quartets

Joseph Rusinko[1,*], Yu Cai[1], Allison Crysler[1], Katherine Thompson[2], Julien Boutte[3], Mark Fishbein[4] and Shannon C. K. Straub[5]

[1]*Department of Mathematics and Computer Science, Hobart and William Smith Colleges, Geneva, NY 14456, USA*
[2]*Department of Statistics, University of Kentucky, Lexington, KY 40536, USA*
[3]*A2Bio2, 13940 Molleges, France*
[4]*Department of Plant Biology, Ecology and Evolution, Oklahoma State University, Stillwater, OK 74078, USA*
[5]*Department of Biology, Hobart and William Smith Colleges, Geneva, NY 14456, USA*
**Correspondence to be sent to: Department of Mathematics and Computer Science, Hobart and William Smith Colleges, Geneva, NY 14456, USA; E-mail: rusinko@hws.edu.*

Received 10 October 2024; reviews returned 13 February 2025; accepted 06 March 2025
Associate Editor: Dr Claudia Solis-Lemus

Abstract.—After collecting large datasets for phylogenomics studies, researchers must decide which genes or samples to include when reconstructing a species tree. Incomplete or unreliable datasets make the empiricist's decision more difficult. Researchers rely on ad hoc strategies to maximize sampling while ensuring sufficient data for accurate inferences. An algorithm called *PickMe* formalizes the sample selection process, assuming that the samples evolved under the tree Multispecies Coalescent Model. We propose a Bayesian framework for selecting samples for species tree analysis. Given a collection of gene trees, we compute a posterior probability for each quartet, describing the likelihood that the species tree displays this topology. From this, we assign individual samples reliability scores computed as the average of a scaled version of the posterior probabilities. *PickMe* uses these weights to recommend which samples to include in a species tree analysis. Analysis of simulated data showed that including the samples suggested by *PickMe* produced species trees closer to the true species trees than both unfiltered datasets and datasets with ad hoc gene occupancy cut-offs applied. To further illustrate the efficacy of this tool, we apply *PickMe* to gene trees generated from target capture data from milkweeds. *PickMe* indicates that more samples could have reliably been included in a previous milkweed phylogenomic analysis than the researchers analyzed without access to a formal methodology for sample selection. Using simulated and empirical data, we also compare *PickMe* to existing sample selection methods. Inclusion of *PickMe* will enhance phylogenomics data analysis pipelines by providing a formal structure for sample selection [*Asclepias*, Apocynaceae, Bayes factor, gene tree, milkweed, phylogenomics, sample selection.]

Accurate phylogenomic analyses rely on careful selection of genes and samples, but this can be challenging in the face of missing data and gene tree estimation error. Empowered by advances in DNA sequencing technologies, phylogenomics researchers have access to large amounts of nuclear gene data generated from whole genome sequences, transcriptomes, or targeted sequencing (Bravo et al., 2019; Lemmon and Lemmon, 2013; McKain et al., 2018). Incorporating a higher number of genes into a phylogenomic analysis can increase phylogenetic accuracy; however, systematists often need larger numbers of samples or sampling of specific taxa to address their questions (Rokas and Carroll, 2005). In practice, the number of genes included in phylogenomics studies and total dataset size has increased more quickly than sampling more taxa or individuals per taxon (Bravo et al., 2019). Systematists must choose which genes and samples to include in a phylogenomic analysis because current methods generate datasets with missing data. Missing data in the form of incomplete taxon or character sampling and their impacts on the accuracy of gene tree and species tree inference have long concerned systematists. Investigations into this issue have yielded contradictory results (Hovmöller et al., 2013; Molloy and Warnow, 2018; Roure et al., 2013; Sayyari et al., 2017a; Wiens and Morrill, 2011). Errors estimating gene trees hinder accurate species tree inference; however, the magnitude of the effect differs based on the specific inference method (Mirarab and Warnow, 2015; Molloy and Warnow, 2018; Sayyari et al., 2017a).

Given the impact of missing data, gene tree estimation error, and other issues affecting phylogenetic analyses on species tree inference, systematists use various strategies to select genes and filter data before analyzing the final dataset. Selection and filtering may happen before data collection, postsequence assembly, multiple sequence alignment, or gene tree inference. Many researchers prefer to sequence or analyze only putatively single-copy genes to avoid orthology-paralogy conflation during data analysis (e.g., de La Harpe et al., 2019; Hime et al., 2021; Li et al., 2021; One Thousand Plant Transcriptomes Initative, 2021; Weitemier et al., 2014; Wielstra et al., 2019). Others sequence gene families and incorporate them into phylogenetic analyses

after careful consideration of orthology (e.g. Laumer et al., 2019; Li et al., 2020; Moore et al., 2018; Smith et al., 2022b; Tice et al., 2021); the choice often depends on data type: target capture sequences versus transcriptomes versus whole genome sequences. After sequence assembly, gene inclusion may depend on the molecular evolution characteristics or phylogenetic informativeness of each gene (e.g. Courcelle et al., 2019; Duchêne et al., 2021; Fernández et al., 2016; Herrando-Moraira et al., 2018; Hosner et al., 2016; Klopfstein et al., 2017; Laumer et al., 2019; Leite et al., 2021; Meiklejohn et al., 2016; Mongiardino Koch, 2021). Often genes are included or excluded based on sample occupancy or aligned matrix column occupancy (e.g. Fernández et al., 2016; Kong et al., 2021; Ontano et al., 2021; Parks et al., 2018a,b). Furthermore, removing fragmentary sequences from alignments, rather than whole genes, before phylogenetic analysis improves gene trees and species trees (Hime et al., 2021; Hosner et al., 2016; Sayyari et al., 2017a). The use of minimum sample occupancy as a criterion for gene inclusion is widespread in animal phylogenomics studies based on ultraconserved element (UCE) targeted sequence data (e.g., Branstetter et al., 2017; Leite et al., 2021; Price et al., 2022; Romiguier et al., 2022). Minimum sample occupancy is often set at 75% based on suggested PHYLUCE (Faircloth, 2015) settings (https://phyluce.readthedocs.io/en/latest/tutorial-one.html#final-data-matrices). After gene tree inference, outlier genes may be detected and discarded (Comte et al., 2023; de Vienne et al., 2012), or gene trees with low bootstrap support for many nodes may be excluded from species tree inference using summary methods (e.g. Boutte et al., 2019; Mongiardino Koch, 2021; Parks et al., 2018a). "Rogue" or "wildcard" samples whose inclusion reduces the resolution of and support for splits in gene trees may be identified and removed (Aberer et al., 2012; Goloboff and Szumik, 2015, and e.g. Foley et al., 2019; Laumer et al., 2019; Li et al., 2021; Salichos and Rokas, 2013). Alternatively, samples identified as outliers by various criteria including differing positions among gene trees or unusually long branch lengths may be removed from individual gene trees (Comte et al., 2023; de Vienne et al., 2012; Mai and Mirarab, 2018, and e.g. Hime et al., 2021; Li et al., 2021). After species tree inference, rogue taxa whose position varies across different analyses are occasionally removed (e.g. One Thousand Plant Transcriptomes Initative, 2019).

Authors of phylogenomics papers rarely describe how they select individual samples for analysis, except when identifying and removing rogue or outlier samples. A percentage of the species sampled in a higher taxonomic group is often referenced as part of a sampling strategy for a phylogenomic study, given that obtaining tissue of all species may be infeasible. A few studies list the subjective ad hoc strategy used to determine which sequenced samples to include. For example, Bagley et al. (2020) explicitly stated that they removed samples with data for fewer than 50% of the 546 loci included in their study. Streicher et al. (2015) explored the effects of including low gene occupancy samples by constructing three datasets that included fewer samples as required gene occupancy increased (all samples, $>$ 2000 UCEs present, $>$ 3000 UCEs present), which were used to create datasets with various minimum sample occupancy criteria. Using multivariate clustering, Gates et al. (2018) developed a method to assess phylogenetic signal for filtering samples from the set of all sequenced individuals for gene and species tree inference. Smith et al. (2022a) identified and filtered samples with higher amounts of missing data and lower contributions to phylogenetic informativeness that contributed to discordance among trees. Because of their occasional explicit description and intuitive logic, we posit that ad hoc minimum gene occupancy cut-offs are often employed in empirical studies but rarely included in the methods sections of manuscripts.

Given that increased sampling improves phylogenetic accuracy (Zwickl and Hillis, 2002), to enable maximum sampling for phylogenomics studies, we introduce *PickMe*. This algorithm recommends which samples to include in a species tree analysis. Using *PickMe* formalizes the sample selection process, replacing ad hoc methodologies requiring the sample to be included in a fixed or minimum percentage of the gene trees.

PickMe is grounded in the hypothesis that the genes evolved under the tree Multispecies Coalescent Model. This model, recently reviewed in Mirarab et al. (2021), captures the discordance between gene and species trees resulting from incomplete lineage sorting. Like many popular algorithms for reconstructing species trees (Chifman and Kubatko, 2014; Mirarab et al., 2014; Zhang et al., 2018), *PickMe* is built on the principle that the distribution of 4 sample subtrees, known as quartets, across gene trees is well-behaved under the tree Multispecies Coalescent Model. Under the tree Multispecies Coalescent Model, quartets displayed by the majority of gene trees consistently estimate the species tree (Allman et al., 2011). Quartets have also played an important role in the analysis of reconstructed species trees (e.g. Larget et al., 2010; Minh et al., 2020; Pease et al., 2018; Sayyari and Mirarab, 2016).

Rather than using quartet data to build species trees, *PickMe* assesses which individual samples can be reliably included in the analysis. To make this assessment, we introduce a Bayesian weight function that assigns a weight to each quartet topology based on the likelihood that the topology appears on the species tree. Each sample is assigned a score by averaging the coalescent weighted quartets across all quartets containing the individual sample. Then, samples are recommended for inclusion (or exclusion) in a species tree analysis based on these scores. While coalescent models have been extended to account for hybridization along a network (Degnan, 2018), *PickMe* does not account for hybridization or introgression.

This study introduces a novel sample selection algorithm *PickMe*. We ground *PickMe* in a Bayesian

framework and explore its effectiveness in simulation studies. We apply *PickMe* to 2 single-copy nuclear gene target capture datasets from flowering plants, revealing that more samples can reliably be included in phylogenomic analyses compared with previous studies, highlighting the value of this methodology in phylogenomic reconstruction pipelines. Furthermore, we compare *PickMe* with existing sample selection methods using simulated and empirical data.

Materials and Methods

Species tree reconstructions play a crucial role in understanding evolutionary relationships and patterns. In this context, the software *PickMe* has been developed to enhance the accuracy and efficiency of species tree reconstructions. The software employs a 3-component process to achieve this goal. First, quartets are scored based on the likelihood of being displayed on a species tree. Subsequently, the average scores associated with quartets containing an individual sample are computed. This average is then interpreted within the framework of Bayesian statistics, providing a robust foundation for selecting samples in the species tree analysis.

In Section 2.1, we develop the mathematical and statistical framework behind *PickMe*. In Section 2.1, we apply *PickMe* to 2 datasets (*Asclepias* and *Iochroma*) as well as conduct a simulation study on the impact of using *PickMe* in the presence of fragmentary sequences.

Mathematical Foundation of PickMe

Coalescent weighted quartets We assume that individual genes evolve under the tree Multispecies Coalescent Model as mathematically formalized in Degnan and Rosenberg (2009). Under this model, the species tree's branch lengths (representing coalescent time) and widths (representing population size) determine the probability distribution of gene trees.

We denote Q_m as the set of all 4-taxon samples on a set of samples $[m] = \{1, 2, \cdots, m\}$. For each 4-taxon sample, there are three possible unrooted binary tree topologies. We refer to these trees as quartets and denote the set of all such quartets $\mathbf{Q}_m$. Thus, it follows that $|\mathbf{Q}_m| = 3|Q_m| = 3\binom{m}{4}$. We use q to denote a generic collection of 4-taxon samples and $\mathbf{q}$ to refer to a quartet.

Under the coalescent model, the quartet tree that occurs with the highest probability matches the topology displayed on the species tree (Allman et al., 2011). Thus, the quartet with the highest probability is called the *dominant quartet*. In *PickMe*, we suppress the model condition that the 2 alternative topologies occur with the same probability, a choice we justify in Supplemental Material A1 which explores the implications of alternative statistical assumptions.

Given that we only have a finite sample of estimated gene trees, we introduce a weight assigned to each quartet topology, which estimates the probability that this quartet is dominant. This weight depends on the number of gene trees where a 4-taxon sample is observed and the number of these resolved as a particular quartet.

Given a binary species tree under the Multispecies Coalescent Model, let p be the probability that a gene tree displays a particular quartet. The weight is the posterior probability that the quartet is dominant ($p > \frac{1}{3}$), with a Uniform(0,1) prior distribution on the probability p, given that the quartet appears on k out of all n estimated gene trees modeled as a Binomial distribution. The probability p is a model parameter, while n and k are quantities observed from the estimated gene trees. Note that n may be smaller than the number of gene trees whose leaves contain q if there are nonbinary gene trees.

Given a quartet $\mathbf{q}$ which occurs on k estimated gene trees and whose associated 4-taxon subset q restricts to a binary quartet on n estimated gene trees, we define the *coalescent weight* of $\mathbf{q}$ as the Bayesian posterior probability the quartet is dominant. More specifically

$$w(\mathbf{q}) = \frac{\int_{\frac{1}{3}}^{1} p^k (1-p)^{n-k}\, dp}{\int_{0}^{1} p^k (1-p)^{n-k}\, dp}.$$

As $w(\mathbf{q})$ is a function of n and k, we alternatively use $w(n,k) = w(\mathbf{q})$ when referencing the general properties of w.

We note that Sayyari and Mirarab (2016) defined a similar measure using a multivariate distribution of the three quartets. Our approach focuses on a single quartet at a time, thus relying on a binomial distribution for the likelihood (rather than a multinomial distribution), and yields a Beta posterior distribution to calculate the probability that the given quartet is dominant.

The following theorem demonstrates that coalescent weights measure confidence that quartet topology matches the species tree.

Theorem 0.1 (proof in Supplemental Material A2)**.** The coalescent weight function $w(n,k)$ has the following properties:

1. For a fixed n if $k_1 > k_2$, then $w(n,k_1) > w(n,k_2)$.
2. For a fixed k if $n_1 > n_2$, then $w(n_1,k) < w(n_2,k)$.
3. $w(n,k) = \begin{cases} < 0.5 & \text{if } \frac{k}{n} < \frac{1}{3}. \\ > 0.5 & \text{if } \frac{k}{n} > \frac{1}{3}. \end{cases}$
4. $\lim_{\lambda \to \infty} w(\lambda n, \lambda k) = \begin{cases} 0 & \text{if } \frac{k}{n} < \frac{1}{3}. \\ 0.5, & \text{if } \frac{k}{n} = \frac{1}{3}. \\ 1 & \text{if } \frac{k}{n} > \frac{1}{3}. \end{cases}$

From coalescent weighted quartets to sample reliability scores The quartet weight $w(\mathbf{q})$ gives confidence that a particular quartet matches a quartet in the species tree. If this weight is near 1, we are confident the quartet topology aligns with the true topology of the species tree. Similarly, if the score is close to 0, we are confident that this

quartet does not match the topology of the species tree. In both cases, we have evidence helpful for placing the four samples in q on the species tree. To make this precise, we define the quartet reliability $r(\mathbf{q}) = 2|w(\mathbf{q}) - \frac{1}{2}|$ which rescales the quartet weight so that higher values indicate more confidence in our understanding of the resolution of a particular quartet on the species tree.

For every sample x, we compute the average reliability score over all quartets containing x. For a collection of quartets, $\mathbf{Q}$, let $\mathbf{Q}(x)$ denote the set of quartets that contain the sample x as a leaf. We can then define the sample reliability by $R(x) = \dfrac{\sum_{\mathbf{q}\in\mathbf{Q}(x)} r(\mathbf{q})}{|\mathbf{Q}(x)|}$.

Statistical interpretation of sample reliability scores In the context of Bayesian Statistics, the Bayes Factor (BF) compares H_1, the hypothesis that the given quartet is dominant, to H_2, the hypothesis that the given quartet is not dominant. If $BF > 1$, there is evidence in favor of the hypothesis of a dominant quartet (H_1).

For an individual quartet, coalescent weight is related to the BF (Kass and Raftery, 1995) quartet by the formula $BF = \frac{w(n,k)}{1-w(n,k)}$, from which it follows that $w(n,k) = \frac{BF}{1+BF}$.

Theorem 0.2 (proof in Supplemental Material A3)**.** For quartet reliability scores $r(\mathbf{q}) > \frac{\kappa-1}{\kappa+1}$, the BF associated to $\mathbf{q}$ is either greater than κ or less than $\frac{1}{\kappa}$.

We define the *BF associated with a sample* to be the maximum of $\left\{\frac{R(x)+1}{R(x)-1}, \frac{R(x)-1}{R(x)+1}\right\}$. It follows that the data associated with an individual sample can be considered reliable at a BF threshold of κ if $R(x)$ exceeds $\frac{\kappa-1}{\kappa+1}$. Selecting a threshold value for the BF cutoff is similar to choosing an α cutoff to test for significance in a hypothesis test. Traditional cutoff values for BF analysis include $3, 10, 30$, or 100 (Kass and Raftery, 1995). We recommend including samples in a species tree reconstruction if the associated BF exceeds the $\kappa = 10$ threshold based on the results of our simulation study in Supplemental Material A4.

PickMe *algorithm details* The software *PickMe* should be used early in a phylogenomic analysis pipeline to help identify the appropriate samples to consider in a species tree analysis. The input for *PickMe* is a collection of estimated gene trees computed using all potential samples. As *PickMe* only uses the estimated gene tree topologies, researchers may use their methodology of choice in the initial gene tree estimation. However, bootstrap trees should not be included in the input as they may skew quartet reliability scores by overcounting the number of observed gene trees (bootstrap values themselves will be ignored by the software).

PickMe outputs a list of samples based on how their associated BFs compare with the standard cutoff thresholds in Bayesian Statistics; categories we denote *very bad*, *bad*, *good*, *very good*, and *exceptional* samples. Table 1 provides the relationship between the *PickMe* classification and the associated BF. We recommend including any good, very good, or exceptional samples in the analysis and excluding any labeled bad or very bad.

PickMe uses estimated gene trees to compute each sample's reliability score. However, sample reliability scores are dependent. In particular, low-scoring samples will reduce the Associated BFs for samples that frequently occur on the same gene trees. We iteratively remove the lowest-scoring samples and recalculate the associated BFs to address the intersample dependency. We repeat until all remaining samples have an associated BF of at least one hundred. The *PickMe* is summarized in Algorithm 1.

Based on the recommendations in the output of *PickMe*, we suggest that researchers run their entire species tree estimation pipeline starting from sequence data using only the suggested samples. As *PickMe* is

Algorithm 1. *PickMe*

Result: List of samples classified according to Table 1
Calculate set of samples: X;
while X *is nonempty* **do**
  Compute coalescent quartet weights for $Qs(X)$;
  Compute sample reliability of each sample in X;
  if $r > .9802$ **then**
    Output all samples in X as Exceptional ;
  else
    Choose sample with smallest reliability score ;
    Output samples with classification type;
    Remove sample from X;
  end
end

TABLE 1. ***PickMe* Categories.** We classify samples based on their associated BF. We recommend samples classified as Good, Very Good, or Exceptional be included in a species tree analysis.

$R(x)$ range	BF range	classification
$0 \le R(x) < 0.5$	$\frac{1}{3} < BF < 3$	Very Bad
$0.5 \le R(x) < 0.8182$	$\frac{1}{10} < BF < \frac{1}{3}$ or $3 < BF < 10$	Bad
$0.8182 \le R(x) < 0.9355$	$\frac{1}{30} < BF < \frac{1}{10}$ or $10 < BF < 30$	Good
$0.9355 \le R(x) < 0.9802$	$\frac{1}{100} < BF < \frac{1}{30}$ or $30 < BF < 100$	Very Good
$0.9802 \le R(x)$	$BF < \frac{1}{100}$ or $100 < BF$	Exceptional

based on the Multispecies Coalescent Model, it has broad applicability in phylogenomic studies regardless of the approach employed for species tree inference. The *PickMe*-identified samples may be analyzed by direct inference of the species tree (e.g., RAxML (Stamatakis, 2014), *Beast (Heled and Drummond, 2009), StarBEAST2 (Ogilvie et al., 2017), and SVDQuartets (Chifman and Kubatko, 2014)) or by indirect inference of the species tree using gene trees (e.g., ASTRAL, (Mirarab et al., 2014)).

Relationship between PickMe *local posterior probabilities* The sample reliability scores implemented in *PickMe* are closely connected to the local posterior probabilities (LPPs) defined in Sayyari and Mirarab (2016) and implemented in ASTRAL versions 4.1 and above. Using slightly different model assumptions, both pieces of software compute a Bayesian posterior probability for each quartet topology. The driving distinction is that LPP assigns scores to the edges of an unrooted tree, and *PickMe* assigns scores to individual samples.

The LPP and *PickMe* scores measure support for different features of a species tree as demonstrated. LPP values are typically displayed as labels for nodes of internal vertices in rooted trees. In Fig. 1, one might be tempted to see the LPP score of 0.582 as an indicator of low support for identifying samples C and D as sisters. However, The LPP value 0.582 reflects support for the partition $\{C\},\{D\}|\{A,B\},\{E,F,G\}$. Given the short branch length between the most recent common ancestor of {A,B}, and the most recent common ancestor of $\{A,B,C,D\}$, it seems more likely that the low LPP value measures uncertainty in assessing the divergence times between most recent common ancestor of $\{A,B\}$ and most recent common ancestor of $\{A,B,C,D\}$. Low LPP support shows a problem with the particular clustering, which need not necessarily be related to the support for the labeled branching. Fortunately, a high LPP value does support the branching associated with the labeled node. Therefore, while LPP and *PickMe* have similar fundamental principles, selecting only well-supported samples identified by *PickMe* does not guarantee high LPP support values on the reconstructed tree.

Analysis of PickMe *using Real and Simulated Datasets*

Impact on species tree reconstruction using PickMe *instead of gene occupancy-based sample selection* Traditional species tree simulation studies assume all samples originate from a single true species tree. Thus, all samples should be considered reasonable. A nuanced simulation study on the impact of *PickMe* on Species Tree analysis would require a robust model of error along the species tree pipeline. We restrict our analysis to errors caused by the presence of fragmentary gene sequences.

Sayyari et al. (2017a) demonstrated that fragmentary gene sequences negatively impact species tree reconstruction. As a result, they proposed a filtering model for each gene in the analysis on which they removed all sites present in fewer than a fixed percentage of samples and all samples with data for fewer than a fixed percentage of the sites. They found the best performance removing sites with greater than 90% gaps and all samples with data for fewer than 33% of the remaining sites. We note the distinction that their method filters samples for individual gene trees, while *PickMe* is designed to remove samples for the entire species tree analysis.

To compare *PickMe* with the practice of occupancy-based sample selection, we used the fragmented sequence data from 10 of their 50 species trees (Sayyari

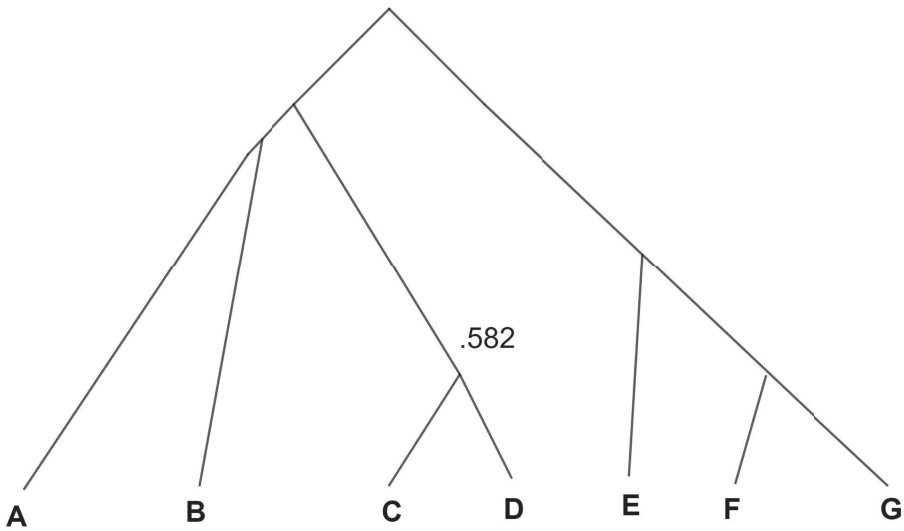

FIGURE 1. **Sample tree** The LPP value may be misleading when plotted above the node rather than the corresponding edge.

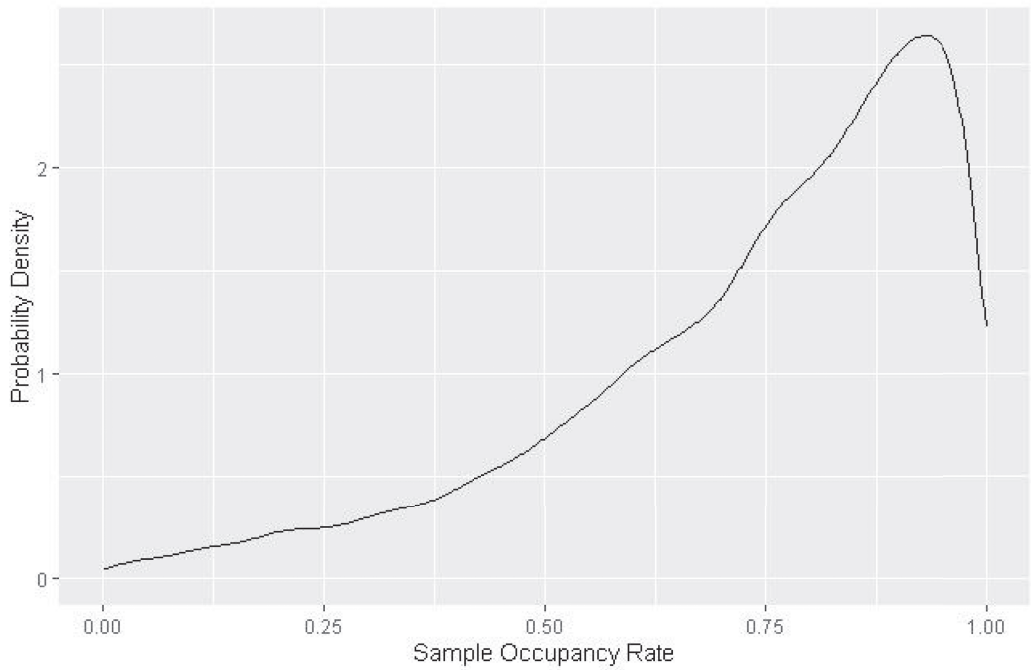

FIGURE 2. **Distribution of sample occupancy rates.** This is a plot of the density function of the distribution under which the sample occupancy rate is drawn. It is intended to mirror the distribution of occupancy rates in biological datasets where most samples have data for most of the genes.

et al., 2017b). Each species tree had 101 samples, including 1 outgroup. We examined a set of 500 generated gene trees and associated fragmented sequences for each species tree. For each of the 10 species trees, we ran 100 trials.

Given that samples in empirical datasets generally do not have full gene occupancy, each sample was assigned an expected occupancy rate. The outgroup was assigned an expected occupancy rate based on a uniform sample on the interval $(0.95, 1)$. The other 100 samples were assigned an expected occupancy rate based on a draw from the distribution shown in Figure 2, which is intended to mirror the distribution of occupancy rates in biological datasets. To construct this distribution, we drew occupancy rates from the beta distribution $B(20, 1.19)$. Only draws in the interval $[.8, 1]$ were used and were then linearly rescaled to the interval $[0, 1]$. Finally, each sample was removed individually from the 500 gene trees if a draw from the uniform random distribution on $[0, 1]$ exceeded its expected occupancy rate.

We used RAxML-NG version 1.2.0 with default parameters to estimate gene trees for these fragmented sequences with missing samples (Kozlov et al., 2019). We then ran *PickMe* on the RAxML-estimated gene trees. For each trial, we dropped all samples identified as Bad or worse by *PickMe* from the gene sequence data, reran RAxML on the reduced set of sequences, and re-estimated the species tree using ASTRAL version 5.7.8 using default parameters (Zhang et al., 2018). Following Sayyari et al. (2017a), we did not realign the sequences after pruning.

For comparison, we removed all samples with gene occupancy rates less than 25%, 50%, and 75%. We ran RAxML-NG version 1.2.0 with default parameters on the reduced set of sequences and the full sequences and estimated the species tree using ASTRAL version 5.7.8.

We computed the Normalized Robinson Foulds distances between the true and estimated species tree using the TreeDist package version 2.6.3 in R (Smith, 2020). An ideal filtering method should have a low species tree estimation error without dropping a large number of samples.

Milkweed (Asclepias) dataset We obtained targeted sequence data for 763 putatively single-copy nuclear loci for samples of 59 North American milkweed species, 3 African outgroup species, *Asclepias physocarpa*, *Asclepias fruticosa*, and *Asclepias fornicata*, and 1 additional outgroup, *Pergularia daemia* using the target enrichment baits of Weitemier et al. (2014) (Supplemental Material A5). Data for 32 of these samples and orthologs from the genome sequence of *Asclepias syriaca* (Weitemier et al., 2019) were included in the analyses of Boutte et al. (2019), and nuclear sequence data for the additional 30 samples were generated using the DNA sequencing and assembly methods described therein. Boutte et al. (2019) had excluded the 30 newly analyzed samples based on an ad hoc minimum gene recovery criterion of 600 genes (79%) with the goal of high gene

occupancy for all samples for species tree analyses. For the analyses conducted here, we masked assembled sequences with Ns for very low read depth (≤ 2 reads) and at heterozygous sites (i.e., intra-individual SNPs). For each gene, we aligned masked sequences using Mafft version 7.245 with default parameters (Katoh and Standley, 2013), and removed sequences with less than 50% of the total alignment length (i.e. Type II missing data; Hosner et al., 2016) to reduce gene tree error following Sayyari et al. (2017a) and Mirarab (2019).

For further analysis, we selected a subset of 703 genes, which had been identified by Boutte et al. (2019) as producing the best-resolved milkweed phylogenies based on bootstrap support across the gene trees. For the complete dataset of 63 species, we first estimated the 703 gene trees using Neighbor-Joining (NJ) on uncorrected and corrected distances using the ape package (Paradis and Schliep, 2018) in R v. 3.5.1 (R Core Team, 2013). To determine whether the gene tree inference method affected the sample selection results, we also used the GTR+Gamma model in RAxML v. 8.2.12; (Stamatakis, 2014) to estimate the initial gene trees. Using these 3 sets of estimated gene trees, we identified the samples to be included in species tree analyses using *PickMe*. If *PickMe* indicated a sample should be excluded for any gene tree set, we excluded it from the downstream analyses. For the set of samples identified as reliable by *PickMe*, we realigned the sequences and then removed small alignments (<100 bp) following Boutte et al. (2019). We then used IQ-Tree v. 1.5.4 (Chernomor et al., 2016; Nguyen et al., 2014) to select the best model of molecular evolution for the retained alignments and inferred the gene tree for each locus using the same parameters as Boutte et al. (2019). Using ASTRAL-II v. 4.10.12 (Mirarab and Warnow, 2015) with default parameters, we inferred a species tree and calculated LPP support (Sayyari and Mirarab, 2016). We calculated gene concordance factors using the method of Minh et al. (2020), implemented in IQ-Tree v. 2.1.2 (Chernomor et al., 2016; Nguyen et al., 2014). For comparison, we repeated the gene and species tree analyses done for the subset of *PickMe* reliable samples for the full dataset using identical methods.

Comparison of sample selection methods on Iochroma *data* The sample selection method *PickMe* plays a similar function to the method for filtering of target capture sequenced individuals proposed in Gates et al. (2018), hereafter referred to as the *Individual Filtering Method*. Both algorithms seek to screen for potentially problematic samples in species tree reconstruction. In the Individual Filtering Method, a matrix of features is computed for all individuals. Then, principal component analysis and machine learning clustering algorithms separate the Good from the Bad samples.

We input the 293 gene trees provided by D. Gates into *PickMe* and compared the suitability classification with those in the published results. In their machine learning cluster analysis, Gates et al. (2018) excluded 4 samples used as outgroups when applying the Individual Filtering Method, though they did appear in the gene trees. Therefore, in our comparison, we assumed that such outgroups were considered good enough for inclusion in their analysis.

Results

Results of simulation measuring the impact of using PickMe *versus gene-occupancy-based selection methods, when estimating species trees* The accuracy of the resulting species tree and the number of samples *PickMe* selected for inclusion varied across the model species trees (Fig. 3).

Use of *PickMe* yielded the highest species tree estimation accuracy across all 1000 trials compared with occupancy cutoffs of 0%, 25%, 50%, or 75% and minimized the number of samples excluded if not all samples were included (Fig. 4). The mean outcomes of this experiment are provided in Table 2.

Analysis of Milkweed Phylogeny Constructed using PickMe-*Recommended Samples*

Sequenced milkweed samples were variable in sequencing depth, target enrichment success (percentage of reads on target), gene recovery (at least some sequence assembled for target genes), and final gene occupancy (nonfragmentary sequences included in matrices used for gene tree inference) (Supplemental Material A5). Among the 30 newly sequenced samples, the three species with the fewest genes recovered (2–4 genes) were removed from all alignments because all of the assembled sequences were fragmentary, and the remaining 27 samples were present on between 3 (0.4%) and 311 (44.2%) gene trees. *PickMe* indicated that 10 of them (gene recovery 199–588 genes/final gene occupancy 64–311) and all samples previously included by Boutte et al. (2019) were suitable for inclusion for species tree inference (Table 3). In general, the unsuitable samples had lower gene recovery and final gene occupancy than suitable samples (Supplemental Material A5). Higher gene recovery did not always translate to higher gene occupancy due to the fragmentary nature of the assembled gene sequences for some samples. Even when only final gene occupancy was considered, not all samples with higher gene numbers were suitable: *A. schaffneri* that was present on 76 (11%) gene trees was not suitable for inclusion when ML and NJ trees based on corrected distances were used for species tree inference, while *Asclepias nivea* that was present on 64 (9%) gene trees was. *Asclepias fruticosa* was present on only 24 (3%) gene trees and yet was able to be included based on ML and uncorrected distance NJ gene trees. The method of gene tree inference only affected which species were selected by *PickMe* for 2 samples (*A. schaffneri* and *A. fruticosa*), and relative confidence in inclusion (e.g., Very Good vs. Exceptional) varied for a few other samples (Table 3).

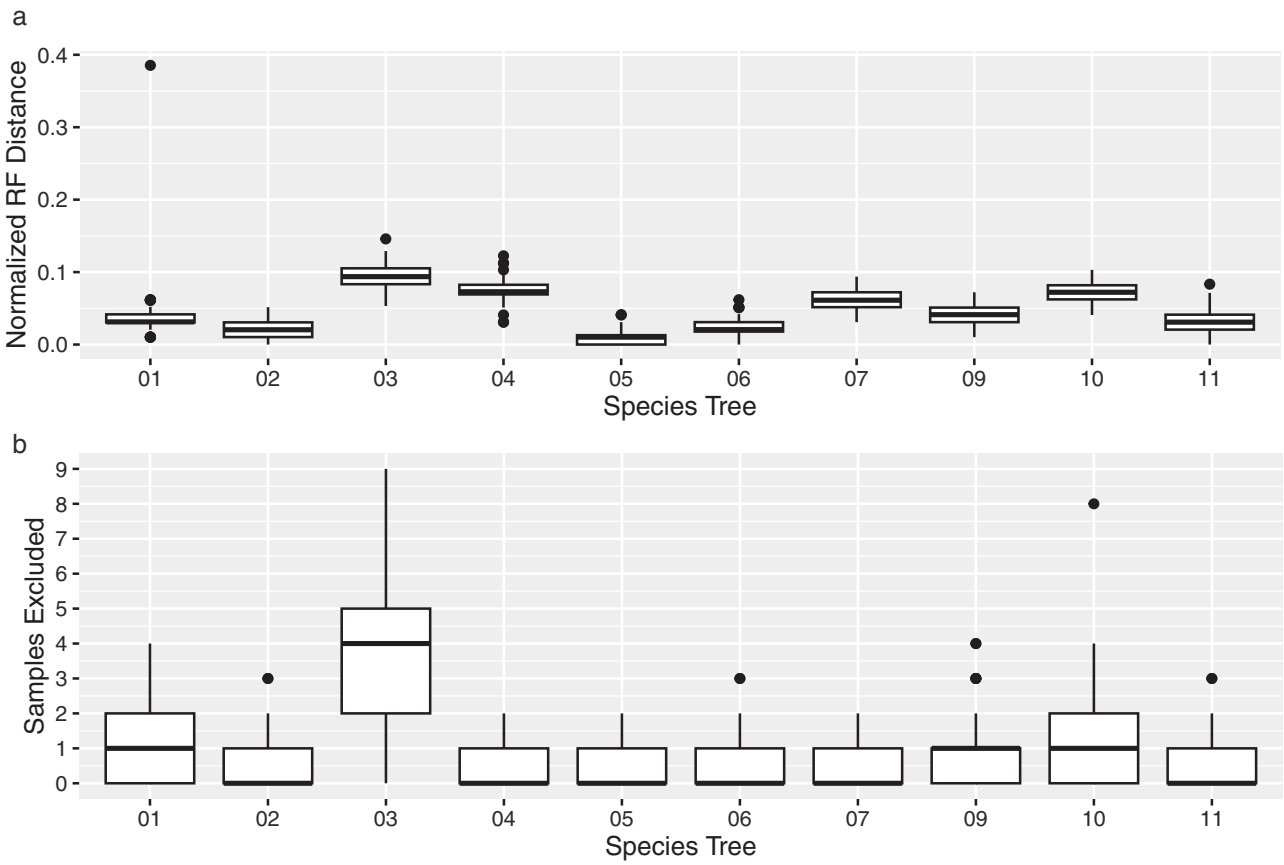

FIGURE 3. **Species tree reconstruction results after sample selection using *PickMe*.** a) Box plot of the normalized Robinson–Foulds distance between the true species tree and the estimated tree after excluding poor-quality samples in 100 trials across 10 species trees. b) Box plot of the number of samples excluded by *PickMe* in these trials. Species tree numbering follows Sayyari et al. (2017b).

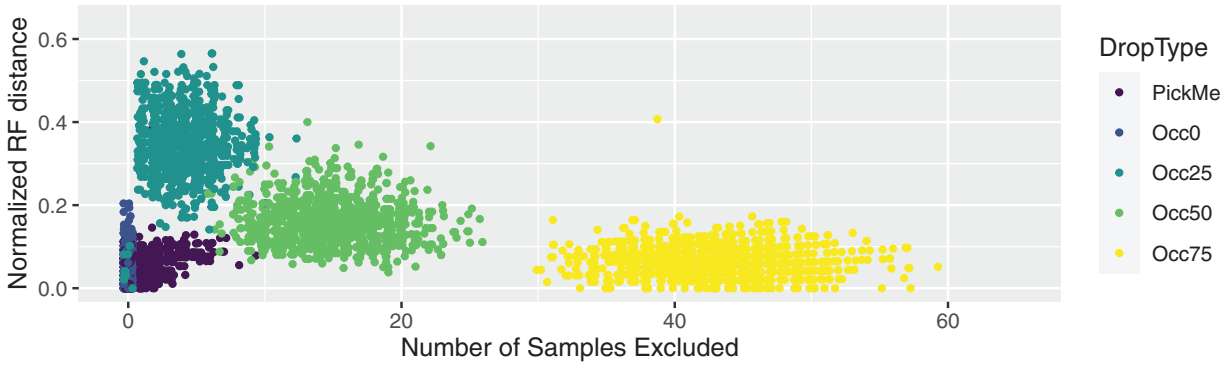

FIGURE 4. **Number of samples excluded vs. species tree accuracy by exclusion method.** For each of the 1000 trials, we plot the Normalized Robinson–Foulds distance between the true and estimated species tree vs. the number of samples excluded by each method. Normalized Robinson–Foulds scores range between 0, (perfect match), to 1 no-matching splits. There were 101 samples prior to exclusion.

The inferred species trees for the samples included using the ad hoc gene recovery cut-off sample set and the set of suitable samples identified using *PickMe* were overall highly similar in topology (Fig. 5). The only difference was the branching order of *A. erosa* and *A. curtissii* when the newly added *A. feayi* resolved as sister to *A. curtissii*. When all samples, both *PickMe* suitable and unsuitable were included in the analysis, the placement of unsuitable samples was evaluated based on which major clade (see Fig. 5), a sample was expected to be a member of based on the recovery of the same major clades in the published milkweed

TABLE 2. **Species tree estimation error and number of samples included by filtering strategy.** The decrease in Normalized Robinson–Foulds error using *PickMe* is significant at the $P < 0.05$ level compared with all occupancy rates. The number of excluded samples is significantly fewer than those using occupancy thresholds of 25% and above and more universally including all samples.

Filtering Strategy	Mean NRF error	Mean samples excluded
PickMe	**.048**	1.0
no filtering	.051	0
Exclude if gene occupancy rate < 25%	0.344	3.9
Exclude if gene occupancy rate < 50%	0.160	14.8
Exclude if gene occupancy rate < 75%	0.065	42.9

TABLE 3. **Milkweed *PickMe* classification and gene occupancy** Classifications computed using gene trees estimated with RAxML. $^{+}$ and $^{\wedge}$ indicate that the sample was classified one category higher when using uncorrected distance or corrected distance Neighbor Joining, respectively. $^{-}$ and $^{\vee}$ indicate that the sample was classified one category lower when using uncorrected distance or corrected distance Neighbor Joining, respectively.

Very Bad Sample	Gene occ.	Bad Sample	Gene occ.	Good Sample	Gene occ.	Very Good Sample	Gene occ.	Exceptional Sample	Gene occ.
A. pumila	0.00	*A. engelmanniana*$^{-\vee}$	0.01	*A. fruticosa*$^{\vee}$	0.03	*A. cutleri*	0.32	*A. angustifolia*$^{-}$	0.26
A. subverticillata	0.00	*A. purpurascens*$^{-\vee}$	0.01	*A. nivea*	0.09	*A. pringlei*	0.45	*Pergularia daemia*$^{\vee}$	0.44
A. aff. *glaucescens*	0.01	*A. contrayerba*	0.03	*A. notha*	0.14	*A. solanoana*	0.67	*A. fornicata*$^{\vee}$	0.66
A. jaliscana	0.01	*A. lynchiana*	0.03	*A. speciosa*	0.21	*A. erosa*$^{-\vee}$	0.97	*A. perennis*	0.67
A. subaphylla	0.01	*A. viridiflora*	0.03	*A. variegata*	0.22	*A. curtissii*	0.99	*A.otarioides*$^{\vee}$	0.71
A. michauxii^{+}	0.02	*A. alticola*	0.04	*A. rubra*	0.27	*A. puberula*	0.99	*A. quinquedentata*$^{-}$	0.77
		Asclepias sp.	0.04	*A. feayi*	0.36	*A. stenophylla*	0.99	*A. syriaca*$^{-}$	0.89
		A. nummularia	0.07	*A. fournieri*$^{+}$	0.38	*A. euphorbiifolia*$^{-\vee}$	1.00	*A. mexicana*	0.92
		A. latifolia	0.08	*A. viridis*	0.62	*A. amplexicaulis*$^{+\wedge}$	1.00	*A. longifolia*	0.96
		A. schaffneri^{+}	0.11	*A. atroviolacea*$^{+}$	0.68			*A. nyctaginifolia*	0.96
				A. scaposa	1.00			*A. albicans*$^{-}$	0.97
								A. oenotheroides	0.98
								A. subulata^{-}	0.98
								A. linaria	0.99
								A. quadrifolia	0.99
								A. tuberosa$^{-\vee}$	0.99
								A. virletii	0.99
								A. aff. *standleyi*	1.00
								A. arenaria	1.00
								A. coulteri	1.00
								A. elata	1.00
								A. emoryi	1.00
								A. hirtella	1.00
								A. physocarpa	1.00

plastome phylogeny Fishbein et al. (2018) and the published nuclear species tree with more limited sampling Boutte et al. (2019). Only 4 of the 17 unsuitable samples were placed outside of their expected major clade (see Supplemental Material A6). Further exploration of the raw data for these four samples (not shown) revealed that two libraries showed evidence of contamination (*A. pumila* and *A. subaphylla*), and one sample, now labeled *Asclepias* sp., was misplaced due to a lab error. There was no detectable issue with the *A. schaffneri* library or data.

PickMe *Recommends More Iochroma Samples than the Individual Filtering Method*

While the Individual Filtering Method recommended excluding 19 of the 50 *Iochroma* samples, all 50 samples met the inclusion criteria of *PickMe*. *PickMe* identified 14 Good, 10 Very Good, and 26 Exceptional samples. Figure 6 shows a general congruence with the analysis of Gates et al. (2018) regarding which samples are better candidates than the others for a species tree analysis. Supplementary Material A7 contains a detailed comparison of the classification results.

DISCUSSION

PickMe Outperforms Both Ad Hoc Gene Occupancy Cut-Offs for Sample Selection and Inclusion of All Samples

Using *PickMe* enabled the inclusion of more samples than pruning samples at the 25%, 50%, or 75% gene occupancy cut-offs while returning more accurate species tree reconstructions. Thus, using this software may empower researchers to analyze more comprehensive datasets than the most commonly used ad hoc occupancy cutoffs without significantly impacting reconstruction accuracy. While our simulation study's error

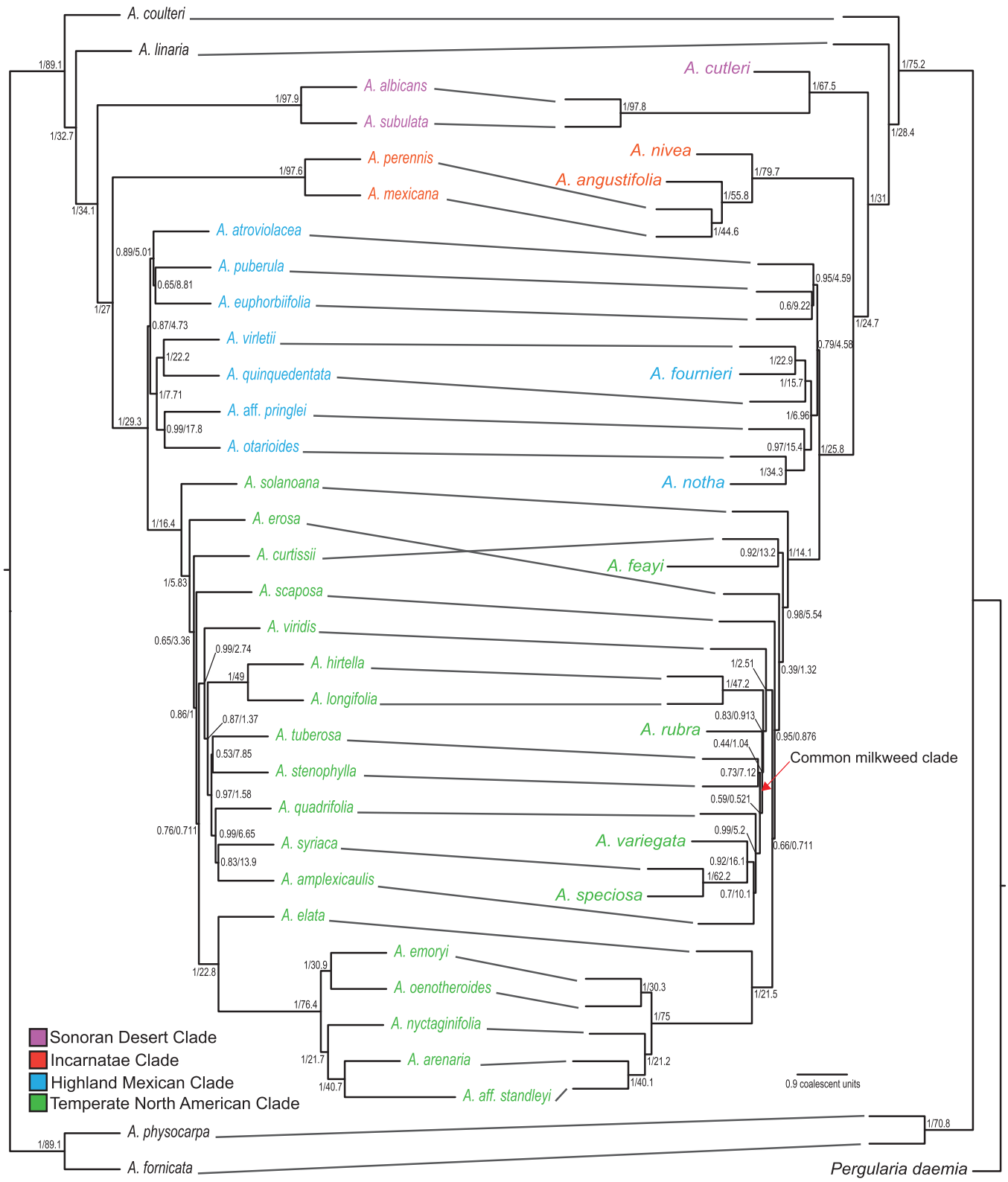

FIGURE 5. **ASTRAL species trees for North American milkweeds (*Asclepias*)**. The tree on the left shows the topology for samples with gene recovery of at least 600 genes. The tree on the right shows the topology with all samples determined to be suitable for inclusion by *PickMe* across all three sets of gene trees (ML, uncorrected, and corrected distance NJ). Numbers near the nodes are LPPs/gene concordance factors.

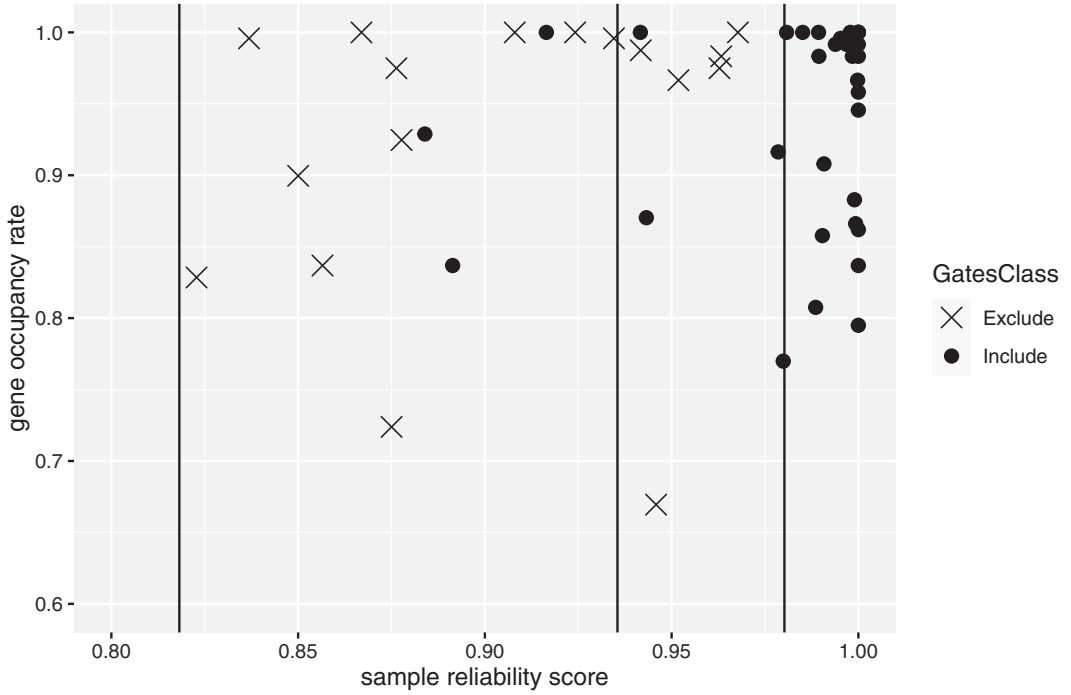

FIGURE 6. **Comparison of sample classification results using *PickMe* and the Individual Filtering Method on *Iochroma* data.** Gene occupancy rate and sample reliability score for 50 *Iochroma* samples. We label samples based on the recommendation of the Individual Filtering Method. Vertical bars delimit Bad/Good, Good/Very Good, from Very Good/Exceptional as classified by *PickMe*.

model is not robust enough to offer definitive advice to practitioners, the results suggest that one should be especially wary about using arbitrary but relatively small gene occupancy cutoffs of 25%, and 50%. In these cases, we found a severe drop in the accuracy of the estimated species trees, in addition to a reduction in the number of samples included. Using the high occupancy threshold of 75%, roughly 43 out of 101 samples were excluded while the species tree accuracy was slightly worse than including all samples.

Using *PickMe* showed a modest but statistically significant improvement in the accuracy of Species Tree reconstruction when measured by normalized Robinson-Foulds distance to the true species tree compared with the inclusion of all samples. However, this may underestimate the practical value of *PickMe*. In practice, fragmentary sequences may harm the alignment process; thus, we recommend rerunning the entire species tree pipeline on the datasets with any samples removed. This simulation does not capture any alignment-driven improvements in gene-tree estimation.

Furthermore, biologists may exclude low-occupancy samples using the intuition that low-occupancy is associated with low-coverage or low-quality. This is not modeled in the simulated data as it is difficult to construct a formal model for low-quality data. However, the κ value simulation indicates that there would be an enhanced benefit of using *PickMe* instead of including all samples if there were more potent sources of error in the gene tree data. In particular, including all samples would not reject any completely random samples. Thus, in practice, *PickMe* may provide a greater improvement over the no filtering method than we found in Table 2. In general, our findings support the assessment of Molloy and Warnow that "the impact of data filtering is fundamentally a question of data quality versus data quantity" Molloy and Warnow (2018).

Implications for the Phylogeny of Asclepias

The analysis conducted by Boutte et al. (2019) was the first to investigate the phylogenomic relationships of *Asclepias* based on data from hundreds of low-copy nuclear genes and indicated that the composition of and relationships among the 4 major clades (Sonoran Desert, Incarnatae, Highland Mexican, Temperate North American) in the nuclear and previously published plastome phylogenies were largely congruent (Fishbein et al., 2011, 2018). Using *PickMe*, Boutte et al. (2019) would have included 9 additional *Asclepias* samples and outgroup *P. daemia*, which had been sequenced with the intention of including them in the initial analysis, but that were later excluded using an ad hoc gene recovery cutoff of 600 genes. In addition to the highly confident placement of the newly added species in the species tree, the expanded sampling resulted in increased support for the positions of some, but not all, species. Gene concordance factors nearly universally decreased, as expected, with the addition of samples that are not well represented on gene trees (Minh et al., 2020).

Inclusion of the new samples identified by *PickMe* increased support for the placement of several species that were unstable in the study of Boutte et al. (2019) with different permutations of the analysis. The placements of *A. atroviolacea* and *A.* aff. *puberula* were especially problematic in the initial analysis because they were alternately resolved as part of the Highland Mexican clade or as sister to the Temperate North American clade. Although the placement of these two species in the Highland Mexican clade agreed with published plastome data (Fishbein et al., 2018), the monophyly of the Highland Mexican clade was not as strongly supported by the nuclear analysis (LPP 0.87) as it was in the plastome phylogeny (100% bootstrap support). After the inclusion of the additional samples identified by *PickMe*, LPP support for the placement of *A. atroviolacea* sister to *A.* aff. *puberula* plus *A. euphorbiifolia* within the Highland Mexican clade increased (LPP 0.89–0.95). However, support for the monophyly of the Highland Mexican clade decreased (LPP 0.87–0.79), even with the newly included *A. fournieri* and *A. notha* confidently placed within it. Furthermore, the new results increased support for other conflicting relationships between the nuclear and plastome topologies that were not unstable in the previous analyses. For example, support for the inclusion of *A. scaposa* in the Temperate North American clade increased (LPP 0.86–0.95), while in the plastome tree, it is not part of, nor closely related to, this clade.

Most of the least well-supported relationships in the species tree occurred in the Temperate North American clade. Adding 3 new species to the common milkweed subclade of the Temperate North American clade (See Fig. 5) generally decreased low support even further. Notable exceptions were the increase in support for the sister relationship of *A. stenophylla* and *A. tuberosa*, which is another conflict with the plastome data, where *A. stenophylla* is not closely related to *A. tuberosa* nor to any other species in this subclade. Likewise, *A. syriaca* and *A. speciosa* are not closely related in the plastome tree, whereas nuclear data indicate that they are well-supported sister species. The latter relationship is supported by numerous morphological similarities and their frequent hybridization (Adams et al., 1987; Fishbein, 2021). Compared with the Boutte et al. (2019) study, adding the new lower gene occupancy species increased our knowledge of species-level relationships in the genus based on nuclear data. It highlighted the difficulty in resolving the species relationships among the most closely related species within the major milkweed clades that, in some cases, are capable of hybridizing (e.g., the common milkweed clade), and which will likely be further complicated with additional sampling of species. It also further underscored the strong incongruence concerning the species relationships within the 4 major clades when nuclear and plastome phylogenies are compared.

The inclusion of samples identified by *PickMe* as unsuitable for comparison yielded results that led to further decreases in support values and placements that were largely in accordance with what might be predicted based on the presumed major clades supported in previous analyses (Boutte et al., 2019; Fishbein et al., 2018). However, given that the true milkweed species tree is unknown, it is impossible to discuss how including the unsuitable samples affected species relationships within those major clades. LPP support values for the Highland Mexican clade (excluding *A. schaffneri*) and the core Temperate North American Clade (i.e., excluding *A. lynchiana* and *A.* aff. *glaucescens*) were slightly decreased (LPP 0.79–0.78 and 1–0.94, respectively) and were likely affected by the conflicting signal present in the 2 oddly-placed samples that were later found to be potentially contaminated. This analysis demonstrated that *PickMe* successfully identified the potentially contaminated samples as unsuitable and which should have been excluded from the final data analysis. In terms of topology, the only major departure from the expected major clade was the placement of *A. schaffneri*, which was strongly supported as part of the Temperate North American clade rather than the Highland Mexican clade. Whether this was due solely to the inclusion of a Bad sample in the analysis will need to be assessed after obtaining data suitable for inclusion in a species tree analysis. Likewise, how relationships among species within the major clades were affected by including Bad samples that generally fell in the expected major clade will need to be reassessed after further nuclear data collection.

Examining LPP Scores through the Lens of PickMe

Through 4 examples drawn from the reconstructed Milkweed tree, we highlight the value of a nuanced approach to interpreting LPP scores.

Example 1: In the Milkweed analysis, *PickMe* classifies the sample *A. angustifolia* as Exceptional despite a gene occupancy rate of only 26%. The edge leading to *A. angustifolia* has LPP = 1, and all adjacent edges continue to have maximal support. Given the relationship between LPP values and the *PickMe* classification, we expect that Exceptional taxa will nearly always be placed in regions of the tree with high LPP support.

Example 2: *Asclepias speciosa* was classified by *PickMe* as Good and had 22% gene occupancy. The node-forward reading of the LPP values on the reconstructed tree would indicate the placement of *A. speciosa* was better supported than that of *A. variegata* (LPP = 1 vs. LPP = 0.92). However, the LPP value of 1 describes the quartet of the form {*A. syriaca*},{*A. speciosa*}|{*A. variegata*},{remaining samples}. This shows strong support not only for the placement of *A. speciosa* as sister to *A. syriaca* but also for *A. variegata* as being sister to the *A. syriaca*−*A. speciosa* clade.

The LPP value of 0.92 on the node to *A. variegata* must therefore be reassessed. It is the support for the quartet {*A. variegata*}, {*A. syriaca*,*A. speciosa*}|{*A. amplexicaulis*,{remaining samples}. In this context, the LPP support of 0.92 is strong given that there was only

a 0.83 support for the {*A. amplexicaulis*},{*A. syriaca*,*A. speciosa*}|{*A. quadrifolia*},{remaining samples} quartet. This suggests that the score of 0.92 may be a remnant of the previously observed difficulty in resolving the relationships among *A. syriaca*, *A. amplexicaulis*, and *A. quadrifolia*.

Example 3: *Asclepias nivea* was classified by *PickMe* as Good while being present on only 9% of gene trees. However, this taxon is placed with LPP = 1, retaining LPP = 1 values around all surrounding edges. In this case, the gene occupancy rate drives the lower *PickMe* classification. For instance, numerous four taxa subsets including *A. nivea* occur on fewer than 10 gene trees. The coalescent quartet scores in these cases can range wildly, even among dominant quartets. In one case, the dominant quartet occurred on 3 out of 7 gene trees, resulting in a reliability score of 0.479, considered Very Bad by *PickMe*. However, in another case, the dominant quartet was displayed by all 6 gene trees, leading to a reliability score of 0.999 that by itself would be considered Exceptional. Thus, with few genes, it may be difficult to have strong support for a dominant quartet whose probability under the coalescent is around $\frac{1}{3}$. In the case of *A. nivea*, the rating of Good reflected the limited information about specific quartets rather than sample misplacement. In conclusion, despite the low coverage *A. nivea* is well supported in the tree.

Example 4: *Asclepias rubra* was classified by *PickMe* as Good while having 27% gene occupancy. Despite occurring on 3 times as many gene trees as *A. nivea*, its coalescent score, while Good, was lower than its well-placed counterpart. The low LPP values on the edges neighboring *A. rubra*, and the relatively low coalescent score, given the occupancy rate, may, at first glance, indicate that this sample should have been excluded from the analysis. However, more carefully analyzing the coalescent quartet scores reveals a different story.

The coalescent quartet weights for *A. rubra* consists largely of very highly supported quartets. Here, over 65% of the quartets that included *A. rubra* had a quartet reliability score of over 0.95, while fewer than 1.4% had a reliability score below 0.05. The majority of low-reliability quartets containing *A. rubra* also included one member of the sister clade {*A. hirtella*, *A. longifolia*}, one member from the common milkweed clade (see Fig. 5), and one of the remaining samples. The low scores reflect the uncertainty regarding the branching order among *A. rubra*, the most recent common ancestor of *A. hirtella*, *A. longifolia*, and the most recent common ancestor of the common milkweed clade. However, a review of the quartets shows that the branching order in the species tree is most prevalent. Given this short internal branch, the coalescent model would require many gene trees to resolve the branching order confidently. We conclude that *PickMe* has correctly determined that there is sufficient evidence to include *A. rubra* in the species tree analysis.

The value of Good samples: The addition of a Good sample is likely to impact LPP in subtle ways. A very narrow range of observations can lead an individual quartet to have a reliability score in the Good range. For example, with 100 gene trees, only dominant quartets that are displayed on exactly 40, 41, or 42 out of the 100 gene trees would have a reliability score in that range. Given that the coalescent taxon score is an average across all quartets containing the sample, it seems that the only way a taxon could be classified as Good is if it had a large number of very well-supported quartets and a much smaller number of very poorly supported quartets. Thus, a Good sample is likely to be well placed on the tree, but an examination of the lowest scoring quartets containing that sample may reveal uncertainties in the species tree.

Comparison of PickMe *with the Individual Filtering Method*

PickMe recommended including all of the *Iochroma* samples, while the Individual Filtering Method of Gates et al. (2018) recommended only around 60% of samples. The discrepancy reflects the methodological differences and the nature of the dataset rather than a dramatic conflict in findings. The Individual Filtering Method separates data into two distinct classes, with those included better suited for species tree analysis than those excluded. In an extreme example, given a uniformly exceptional (or terrible) dataset, the Individual Filtering Method would divide those into two groups. Alternatively, *PickMe* uses fixed, predetermined thresholds rather than localized clustering. Given that the Individual Filtering Method included all *Iochroma* samples *PickMe* identified as Exceptional, our methodology generally supports their classification but allows the inclusion of a larger proportion of the data in the species tree analysis with high confidence. Furthermore, the lowest gene occupancy in the *Iochroma* data was high (70%), suggesting that lower occupancy samples may have been excluded before their analysis. If so, we expect an increase in the overlap between the two methods' classifications if more of the lower occupancy samples were included. Both approaches highlight the importance of having an objective, rather than ad-hoc, methodology for this component of a phylogenomic analysis pipeline.

Limitations of the PickMe Sample Selection Procedure

There are several limitations of the sample selection procedure that users should be aware of. The weight formula in *PickMe* does not require the nondominant quartets to occur with the same frequency. As a result, the algorithm may recommend samples that are hybrids under the Multispecies Coalescent Model. Thus, *PickMe* should not be used as a test for hybridization. Additionally, *PickMe* is designed for input including only a single individual per species. Future software versions could

be extended to allow multiple individuals per species, although it may be difficult to determine when there is sufficient support across a collection of individuals to justify analyzing the species itself in the final analysis. Analysis of the sample selection problem would also benefit from a biologically meaningful error model that systematically generates both good and bad samples for further simulation studies.

Conclusions

Given the impacts of missing data on phylogenetic inference, systematists seeking to limit missing data in their phylogenomic studies have applied numerous data filtering strategies, including excluding samples with low gene occupancy. The *PickMe* procedure presented here provides an objective, repeatable, and easy-to-implement process by which samples with data from multiple genes can be selected for inclusion in phylogenomic datasets. Researchers can confidently increase sampling, thus reducing the costs of additional data generation and increasing the utility of the phylogenies produced for downstream applications that benefit from more complete sampling, such as ancestral state reconstruction and biogeographical analysis.

Acknowledgments

We thank Aaron Liston, Rich Cronn, and Kevin Weitemier for assistance collecting the milkweed data, and Jesse Schafer of the OSU High Performance Computing Center for computational support. We also thank David Eck for data access and management support at Hobart and William Smith Colleges. We thank Dan Gates and Stacey Smith for access to the trees used in their work on *Iochroma*.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.3r2280ggv.

Funding

This work was supported by the National Science Foundation (DMS 1616186 to J.R. and DEB 1457510/1457473 to M.F. and S.C.K.S.) The Oklahoma State University High Performance Computing Center was supported in part through the National Science Foundation grant OCI–1126330. J.R. and K.T. were supported by the National Science Foundation under Grant No. DMS-1929284 while in residence at the Institute for Computational and Experimental Research in Mathematics in Providence, RI, during the Theory Methods, and Applications of Quantitative Phylogenomics program.

Data availability

The raw sequence data are deposited at NCBI under Bioproject PRJNA421764 (BioSample accessions SAMN00779514, SAMN08153045, SAMN08153053, SAMN08153059, SAMN08153064, and SAMN46487195 - SAMN46487218).

PickMe is implemented in the Julia programming language and may be installed directly from the Julia package repository through the package *PhyloPickMe*. *PickMe* is also available at http://github.com/jrusinko/PhyloPickMe.jl.

All the original data and scripts necessary to reproduce the analyses reported in this study can be accessed through the Dryad Digital Repository: Data for: PickMe: sample selection for species tree reconstruction using coalescent weighted quartets submitted with DOI http://doi.org/10.5061/dryad.3r2280ggv. Supplemental Appendices A1 through A7 and project data are also available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.3r2280ggv.

References

Aberer A. J., Krompass D., Stamatakis A. 2012. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Syst. Biol. 62:162–166.

Adams R. P., Tomb A., Price S. 1987. Investigation of hybridization between *Asclepias speciosa* and *A. syriaca* using alkanes, fatty acids and triterpenoids. Biochem. Syst. Ecol. 15:395–399.

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

Bagley J. C., Uribe-Convers S., Carlsen M. M., Muchhala N. 2020. Utility of targeted sequence capture for phylogenomics in rapid, recent angiosperm radiations: neotropical Burmeistera bellflowers as a case study. Mol. Phylogenet. Evol. 152:106769.

Boutte J., Fishbein M., Liston A., Straub S. C. K. 2019. NGS-Indel Coder: a pipeline to code indel characters in phylogenomic data with an example of its application in milkweeds (*Asclepias*). Mol. Phylogenet. Evol. 139:106534.

Branstetter M. G., Danforth B. N., Pitts J. P., Faircloth B. C., Ward P. S., Buffington M. L., Gates M. W., Kula R. R., Brady S. G. 2017. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. Curr. Biol. 27:1019–1025.

Bravo G. A., Antonelli A., Bacon C. D., Bartoszek K., Blom M. P., Huynh S., Jones G., Knowles L. L., Lamichhaney S., Marcussen T., Morlon H., Nakhleh L. K., Oxelman B., Pfeil B., Schliep A., Wahlberg N., Werneck F. P., Wiedenhoeft J., Willows-Munro S., Edwards S. V. 2019. Embracing heterogeneity: coalescing the tree of life and the future of phylogenomics. PeerJ 7:e6399.

Chernomor O., von Haeseler A., Minh B. Q. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. Syst. Biol. 65:997–1008.

Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinform. 30:3317–3324.

Comte A., Tricou T., Tannier E., Joseph J., Siberchicot A., Penel S., Allio R., Delsuc F., Dray S., de Vienne D. M. 2023. PhylteR: efficient identification of outlier sequences in phylogenomic datasets. Mol. Biol. Evol. 40:msad234.

Courcelle M., Tilak M.-K., Leite Y. L., Douzery E. J., Fabre P.-H. 2019. Digging for the spiny rat and hutia phylogeny using a gene capture approach, with the description of a new mammal subfamily. Mol. Phylogenet. Evol. 136:241–253.

de La Harpe M., Hess J., Loiseau O., Salamin N., Lexer C., Paris M. 2019. A dedicated target capture approach reveals variable genetic markers across micro-and macro-evolutionary time scales in palms. Mol. Ecol. Res. 19:221–234.

de Vienne, D. M., S. Ollier, and G. Aguileta. 2012. Phylo-MCOA: A fast and efficient method to detect outlier genes and species in Phylogenomics Using Multiple Co-inertia Analysis. Molecular Biology and Evolution 29:1587–1598.
Degnan J. H. 2018. Modeling hybridization under the network multispecies coalescent. Syst. Biol. 67:786–799.
Degnan, J. H., Rosenberg, N. A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology and Evolution 24:332–340.
Duchêne D. A., Mather N., Van Der Wal C., Ho S. Y. W. 2021. Excluding loci with substitution saturation improves inferences from phylogenomic data. Syst. Biol. 71:676–689.
Faircloth B. C. 2015. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinform. 32:786–788.
Fernández R., Edgecombe G. D., Giribet G. 2016. Exploring phylogenetic relationships within Myriapoda and the effects of matrix composition and occupancy on phylogenomic reconstruction. Syst. Biol. 65:871–889.
Fishbein M. 2021. *Asclepias*. In: Flora of North America (Flora of North America Editorial Committee, eds. 1993+., ed.) vol. 14. Oxford University Press, Oxford and New York (in press).
Fishbein M., Chuba D., Ellison C., Mason-Gamer R. J., Lynch S. P. 2011. Phylogenetic relationships of *Asclepias* (Apocynaceae) inferred from non-coding chloroplast DNA sequences. Syst. Bot. 36:1008–1023.
Fishbein M., Straub S. C., Boutte J., Hansen K., Cronn R. C., Liston A. 2018. Evolution at the tips: *Asclepias* phylogenomics and new perspectives on leaf surfaces. Am. J. Bot. 105:514–524.
Foley S., Lüddecke T., Cheng D.-Q., Krehenwinkel H., Künzel S., Longhorn S. J., Wendt I., von Wirth V., Tänzler R., Vences M., Piel W. H. 2019. Tarantula phylogenomics: a robust phylogeny of deep theraphosid clades inferred from transcriptome data sheds light on the prickly issue of urticating setae evolution. Mol. Phylogenet. Evol. 140:106573.
Gates D. J., Pilson D., Smith S. D. 2018. Filtering of target sequence capture individuals facilitates species tree construction in the plant subtribe Iochrominae (Solanaceae). Mol. Phylogenet. Evol. 123:26–34.
Goloboff P. A., Szumik C. A. 2015. Identifying unstable taxa: efficient implementation of triplet-based measures of stability, and comparison with Phyutility and RogueNaRok. Mol. Phylogenet. Evol. 88:93–104.
Heled J., Drummond A. J. 2009. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.
Herrando-Moraira S., Calleja J. A., Carnicero P., Fujikawa K., Galbany-Casals M., Garcia-Jacas N., Im H.-T., Kim S.-C., Liu J.-Q., Lopez-Alvarado J., López-Pujol J., Mandel J.R., Massó S., Mehregan I., Montes-Moreno N., Pyak E., Roquet C., Sáez L., Sennikov A., Susanna A., Vilatersana R. 2018. Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe Cardueae (Compositae). Mol. Phylogenet. Evol. 128:69–87.
Hime P. M., Lemmon A. R., Lemmon E. C. M., Prendini E., Brown J. M., Thomson R. C., Kratovil J. D., Noonan B. P., Pyron R. A., Peloso P. L. V., Kortyna M. L., Keogh J. S., Donnellan S. C., Mueller R. L., Raxworthy C. J., Kunte K., Ron S. R., Das S., Gaitonde N., Green D. M., Labisko J., Che J., Weisrock D. W. 2021. Phylogenomics reveals ancient gene tree discordance in the amphibian tree of life. Syst. Biol. 70:49–66.
Hosner P. A., Faircloth B. C., Glenn T. C., Braun E. L., Kimball R. T. 2016. Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). Mol. Biol. Evol. 33:1110–1125.
Hovmöller R., Knowles L. L., Kubatko L. S. 2013. Effects of missing data on species tree estimation under the coalescent. Mol. Phylogenet. Evol. 69:1057–1062.
Kass R. E., Raftery A. E. 1995. Bayes factors. J. Am. Stat. Assoc. 90:773–795.
Katoh K., Standley D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.
Klopfstein S., Massingham T., Goldman N. 2017. More on the best evolutionary rate for phylogenetic analysis. Syst. Biol. 66:769–785.
Kong, H., F. L. Condamine, L. Yang, A. J. Harris, C. Feng, F. Wen, and M. Kang. 2021. Phylogenomic and macroevolutionary evidence for an explosive radiation of a plant genus in the miocene. Systematic Biology 71:589–609.
Kozlov A. M., Darriba D., Flouri T., Morel B., Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinform. 35:4453–4455.
Larget B. R., Kotha S. K., Dewey C. N., Ané C. 2010. BUCKy: gene tree/species tree reconciliation with bayesian concordance analysis. Bioinform. 26:2910–2911.
Laumer C. E., Fernández R., Lemer S., Combosch D., Kocot K. M., Riesgo A., Andrade S. C., Sterrer W., Sørensen M. V., Giribet G. 2019. Revisiting metazoan phylogeny with genomic sampling of all phyla. Proc. Royal Soc. B 286:20190831.
Leite R. N., Kimball R. T., Braun E. L., Derryberry E. P., Hosner P. A., Derryberry G. E., Anciaes M., McKay J. S., Aleixo A., Ribas C. C., Brumfield R.T., Cracraft J. 2021. Phylogenomics of manakins (Aves: Pipridae) using alternative locus filtering strategies based on informativeness. Mol. Phylogenet. Evol. 155:107013.
Lemmon E. M., Lemmon A. R. 2013. High-throughput genomic data in systematics and phylogenetics. Ann. Rev. Ecol. Evol. Syst. 44:99–121.
Li J., Lemer S., Kirkendale L., Bieler R., Cavanaugh C., Giribet G. 2020. Shedding light: a phylotranscriptomic perspective illuminates the origin of photosymbiosis in marine bivalves. BMC Evol. Biol. 20:1–15.
Li Y., Steenwyk J. L., Chang Y., Wang Y., James T. Y., Stajich J. E., Spatafora J. W., Groenewald M., Dunn C. W., Hittinger C. T., Shen X.-X., Rokas A. 2021. A genome-scale phylogeny of the kingdom Fungi. Curr. Biol. 31:1653–1665.e5.
Mai U., Mirarab S. 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics 19:272.
McKain M. R., Johnson M. G., Uribe-Convers S., Eaton D., Yang Y. 2018. Practical considerations for plant phylogenomics. Appl. Plant Sci. 6:e1038.
Meiklejohn K. A., Faircloth B. C., Glenn T. C., Kimball R. T., Braun E. L. 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements: evidence for a bias in some multispecies coalescent methods. Syst. Biol. 65:612–627.
Minh B. Q., Hahn M. W., Lanfear R. 2020. New methods to calculate concordance factors for phylogenomic datasets. Mol. Biol. Evol. 37:2727–2733.
Mirarab S. 2019. Species tree estimation using ASTRAL: practical considerations. Preprint. arXiv. https://doi.org/10.48550/arXiv.1904.03826.
Mirarab S., Nakhleh L., Warnow T. 2021. Multispecies coalescent: theory and applications in phylogenetics. Ann. Rev. Ecol. Evol. Syst. 52:247–268.
Mirarab S., Reaz R., Bayzid M. S., Zimmermann T., Swenson M. S., Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinform. 30:i541–i548.
Mirarab S., Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinform. 31:i44–i52.
Molloy E. K., Warnow T. 2018. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67:285–303.
Mongiardino Koch N. 2021. Phylogenomic subsampling and the search for phylogenetically reliable loci. Mol. Biol. Evol. 38:4025–4038.
Moore A. J., Vos J. M. D., Hancock L. P., Goolsby E., Edwards E. J. 2018. Targeted enrichment of large gene families for phylogenetic inference: phylogeny and molecular evolution of photosynthesis genes in the Portullugo clade (Caryophyllales). Syst. Biol. 67: 367–383.
Nguyen L.-T., Schmidt H. A., von Haeseler A., Minh B. Q. 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32: 268–274.
Ogilvie H. A., Bouckaert R. R., Drummond A. J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. 34:2101–2114.

One Thousand Plant Transcriptomes Initative. One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574:679–685.

Ontano A. Z., Gainett G., Aharon S., Ballesteros J. A., Benavides L. R., Corbett K. F., Gavish-Regev E., Harvey M. S., Monsma S., Santibáñez-López C. E., Setton E. V. W., Zehms J. T., Zeh J. A., Zeh D. W., Sharma P. P. 2021. Taxonomic sampling and rare genomic changes overcome long-branch attraction in the phylogenetic placement of pseudoscorpions. Mol. Biol. Evol. 38: 2446–2467.

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

Paradis E., Schliep K. 2018. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinform. 35: 526–528.

Parks M. B., Nakov T., Ruck E. C., Wickett N. J., Alverson A. J. 2018a. Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). Am. J. Bot. 105:330–347.

Parks M. B., Wickett N. J., Alverson A. J. 2018b. Signal, uncertainty, and conflict in phylogenomic data for a diverse lineage of microbial eukaryotes (diatoms, Bacillariophyta). Mol. Biol. Evol. 35: 80–93.

Pease J. B., Brown J. W., Walker J. F., Hinchliff C. E., Smith S. A. 2018. Quartet sampling distinguishes lack of support from conflicting support in the green plant tree of life. Am. J. Bot. 105: 385–403.

Price S. L., Blanchard B. D., Powell S., Blaimer B. B., Moreau C. S. 2022. Phylogenomics and fossil data inform the systematics and geographic range evolution of a diverse neotropical ant lineage. Insect Syst. Diver. 6:9.

R Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria.

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

Rokas A., Carroll S. B. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22:1337–1344.

Romiguier J., Borowiec M. L., Weyna A., Helleu Q., Loire E., La Mendola C., Rabeling C., Fisher B. L., Ward P. S., Keller L. 2022. Ant phylogenomics reveals a natural selection hotspot preceding the origin of complex eusociality. Curr. Biol. 32:2942–2947.e4.

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

Roure B., Baurain D., Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30:197–214.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Sayyari E., Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33:1654–1668.

Sayyari E., Whitfield J. B., Mirarab S. 2017a. Fragmentary gene sequences negatively impact gene tree and species tree reconstruction. Mol. Biol. Evol. 34:3279–3291.

Sayyari E., Whitfield J. B., Mirarab S. 2017b. Fragmentary gene sequences negatively impact gene tree and species tree reconstruction [dataset].

Smith B. T., Merwin J., Provost K. L., Thom G., Brumfield R. T., Ferreira M., Mauck I., William M., Moyle R. G., Wright T. F., Joseph L. 2022a. Phylogenomic analysis of the parrots of the world sistinguishes artifactual from biological sources of gene tree discordance. Syst. Biol. 72:228–241.

Smith M. L., Vanderpool D., Hahn M. W. 2022b. Using all gene families vastly expands data available for phylogenomic inference. Mol. Biol. Evol. 39:msac112.

Smith M. R. 2020. TreeDist: distances between phylogenetic trees. R package version 2.6.3.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinform. 30:1312–1313.

Streicher J. W., Schulte I., James A., Wiens J. J. 2015. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. Syst. Biol. 65:128–145.

Tice A. K., Žihala D., Pánek T., Jones R. E., Salomaki E. D., Nenarokov S., Burki F., Eliáš M., Eme L., Roger A. J., Rokas A., Shen X.-X., Strassert J. F. H., Kolísko M., Brown M. W. 2021. PhyloFisher: a phylogenomic package for resolving eukaryotic relationships. PLOS Biol. 19:1–22.

Weitemier K., Straub S. C., Cronn R. C., Fishbein M., Schmickl R., McDonnell A., Liston A. 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. Appl. Plant Sci. 2:1400042.

Weitemier K., Straub S. C., Fishbein M., Bailey C. D., Cronn R. C., Liston A. 2019. A draft genome and transcriptome of common milkweed (*Asclepias syriaca*) as resources for evolutionary, ecological, and molecular studies in milkweeds and apocynaceae. PeerJ 7:e7649.

Wielstra B., McCartney-Melstad E., Arntzen J., Butlin R., Shaffer H. 2019. Phylogenomics of the adaptive radiation of *Triturus* newts supports gradual ecological niche expansion towards an incrementally aquatic lifestyle. Mol. Phylogenet. Evol. 133:120–127.

Wiens J. J., Morrill M. C. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60:719–731.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinform. 19:15–30.

Zwickl D. J., Hillis D. M. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.