Contents

List of Figures

List of Tables

S1 Simulation Design

The data were generated using SimPhy (version 1.0.2) [1] under the DLCoal model [2]. We generated the default dataset with the following command:

```
simphy -sl f:100 -rs 10 -rl f:1000 -rg 1 -sb f:0.000000005 \
    -sd f:0 -st ln:21.25,0.2 -so f:1 -si f:1 -sp f:50000000 \
    -su ln:-21.9,0.1 -hh f:1 -hs ln:1.5,1 \
    -hl ln:1.551533,0.6931472 -hg ln:1.5,1 -cs 8472 -v 3 \
    -o default -ot 0 -op 1 -lb f:0.0000000005 \
    -ld f:0.0000000005 -lt f:0
```

The parameters that do not directly correspond to parameters mentioned in the "material and methods" section were chosen based on the data simulated for the original ASTRAL-Pro study [3]. Some of these parameters (`-hh`, `-hs`, `-hl`, and `-hg`) are used to deviate from the strict molecular clock. We used INDELible [4] to simulate sequences down the true gene trees produced by SimPhy, which were provided with branch lengths. We downloaded the INDELible executable from the link: `http://abacus.gene.ucl.ac.uk/software/indelible/download.php`. INDELible takes as input a control file containing inputs and all possible configurations and outputs the sequences in PHYLIP format. In the control file, we chose the Generalised Time Reversible (GTR) model [5] for the nucleotide substitution model. The GTR parameters for the substitution rate matrix were generated from a Dirichlet distribution, where the parameters of the distribution were chosen based on the data simulated in the FastMulRFS study [6]. Finally, we used the continuous Gamma distribution as the rates-across-sites model, and the shape parameter of the Gamma distribution has been taken from a lognormal distribution. We used the parameters for the lognormal distribution from the FastMulRFS study [6].

Table S1: INDELible Parameters

Parameters	Values
Base Frequencies	Dirichlet(T=113.48869,C=69.02545,A=78.66144,G=9983793)
Transition Rate	Dirichlet(CT=12.776722,AT=20.869581,GT=5.647810, AC=9.863668,GC=30.679899,AG=3.199725)
Gamma rate variation	Log-normal(μ-0.470703916, σ=0.348667224)

We generated one sequence per leaf and automated the creation of the control files and the running of INDELible with scripts `set_indelible_params.py` and `run_indelible.py`, both written by Erin Molloy and available at: `https://databank.illinois.edu/datasets/IDB-5721322` (in tools.zip). The commands (with Dirichlet distribution parameters) used in the study are:

```
python set_indelible_params.py \
        -n $number_of_total_gene_trees \
        -f 113.48869 69.02545 78.66144 99.83793 \
        -r 12.776722 20.869581 5.647810 9.863668 30.679899 3.199725 \
        -a -0.470703916 0.348667224 \
         -l $sequence_length \
         -o indelible-parameters.csv

python run_indelible.py \
         -x path/to/indelible/executable \
         -s 1 \
         -e $number_of_total_gene_trees \
         -p indelible-parameters.csv \
         -t true_gene_trees.trees \
         -o output/directory/for/estimated/gene/trees
```

To generate datasets with missing taxa, we ran the script `missing.py` (available from `https://github.com/roddur/Generate-missing-taxa-dataset`). The script determines the clades in the species tree containing at least 20% of all the leaves in the tree; then, for each of the selected gene-family trees (the probability that a gene-family will be selected is 0.95), it randomly (uniformly) picks one of these clades and deletes all the leaves from that gene family tree not in the selected clade. We did not delete species from gene trees containing fewer than 20% of the total species.

```
python missing.py -g genes.trees > genes_missing.trees
```

S2 Methods, Commands, and Version Numbers

All of these commands were run on the Campus Cluster of the University of Illinois Urbana-Champaign or on Tallis, which has 1 node with 16 threads; further information about the Campus Cluster hardware can be found here: `https://campuscluster.illinois.edu/about/system-info/hardware/`.

S2.1 ASTRAL-Pro

We ran ASTRAL-Pro (version 1.1.5) on our datasets with the following command:

```
 java -jar -D"java.library.path=lib" astral.1.1.5.jar \
     -i genes-mult.trees -o species.tree
```

When running on the 1001 species conditions we included the argument `-t 0`, which disables the computation of branch support. We did this as ASTRAL-Pro would often crash on these model conditions otherwise. Additionally, we computed localPP branch support with the following command:

```
java -D"java.library.path=lib" astral.1.1.5.jar \
    -q species.tree -i gene-mult.trees -o species-scored.tree
```

S2.2 ASTRID-multi

We also ran ASTRID-multi (version 2.2.1) on our simulated datasets as follows:

```
ASTRID -i gene.trees -a s2g-map.txt -o species.tree -u -n -s
```

S2.3 DISCO decomposition and analyses

When running DISCO with ASTRID or ASTRAL, we used the following command:

```
python disco.py -i genes-mult.trees
```

Then we either ran ASTRID (version 2.2.1) or ASTRAL (version 5.15.3 and version 5.15.4) as:

```
ASTRID -i genes-mult-decomp.trees -o species.tree -u -n -s
```

and

```
java -D"java.library.path=lib/" -jar astral.5.15.3.jar \
    -i genes-mult-decomp.trees -o species.tree -C
```

respectively. The ASTRID `-u -n -s` command tells ASTRID what distance methods to use (more information in ASTRID's readme), and `-C` disables using GPU for ASTRAL (we did not have a GPU available).

When running DISCO with concatenation analysis (e.g., IQ-TREE), we used the following command:

```
python concat.py -i gene.trees -o concat_seq.phy \
    -aln align_directory -d d
```

The script used to decompose the gene trees and create the concatenated alignment (`concat.py`) is available here: `https://gist.github.com/RuneBlaze/249fff484c339e04dcde9c79f9ff52fd`. A new improved script (`ca_disco.py`) is available at the DISCO GitHub site.

S2.4 FastMulRFS

FastMulRFS (version 3) works in two steps. First, we preprocess input multi-label gene trees to generate singly-labeled gene trees. We downloaded the Python script for FastMulRFS from the link: https://github.com/ekmolloy/fastmulrfs. We ran this step with the following command.

```
python  preprocess_multrees_v3.py -i gene-mult.trees \
    -o gene-preprocessed.trees
```

FastMulRFS uses FastRFS [7] to compute the species tree, and FastRFS uses ASTRAL [8] (version 5.7.1) to calculate its constraint set. The executable for FastRFS version 1.0 is available at https://github.com/pranjalv123/FastRFS/releases.

```
FastRFS -i gene-preprocessed.trees -o species.tree
```

FastRFS by default uses SIESTA [9], which returns the strict consensus, majority consensus, and greedy consensus trees of the set of all optimal trees. However, we reported the results for a single (binary) tree returned that has the best RFS score.

S2.5 FastTree

We used FastTree (version 2.1.11) (see http://www.microbesonline.org/fasttree/) to estimate gene trees from the sequences generated using INDELible, using command:

```
FastTree -nt -gtr gene.phy > gene.tree
```

S2.6 IQ-TREE

For our concatenation analysis we ran IQ-TREE (version 1.6.12), using the following command:

```
iqtree -s concat_seq.phy -m GTR+G
```

S2.7 Maximum Inclusion (MI) decomposition

We used the MI implementation provided by [10]. In order to use their implementation in the experiments, their script was modified to reflect that the taxon-individual separator is "_" instead of "@" in the DISCO datasets. The modified version of the script is available at https://github.com/RuneBlaze/phylogenomic_dataset_construction.

The MI script was run using the following command:

```
python2.7 scripts/prune_paralogs_MI.py gene-mult "" inf inf 4 outdir
```

The input gene family trees must be individually put into different files under the same directory gene-mult.

S2.8 Robinson-Foulds errors, MGTE, and AD

We measured the normalized Robinson-Foulds distance between the estimated species trees and the true species trees, using a script written by Erin Molloy (see the FastMulRFS GitHub repository at `https://github.com/ekmolloy/fastmulrfs`). These values were normalized by dividing by $n-3$, where n is the number of leaves in the tree, to produce RF error rates.

MGTE and AD were calculated using the same scripts written by Erin Molloy. For MGTE we computed the average Robinson-Foulds (RF) distance between the true gene family trees and estimated gene family trees; the script works only with single copy trees, however the trees outputted by Simphy have each copy of a gene labeled uniquely, allowing us to use this metric. For AD, we used the same script, this time between the locus trees and true gene family trees outputted by SimPhy.

S2.9 Running Time and Memory

We measure the total running time as well as the peak memory usage. We measure the pipeline starting before the gene trees are decomposed until after the species tree is returned by recording the time with `date -u +%s.%N`, then taking the difference. Peak memory is recorded for each step using:

```
/usr/bin/time -f "%M" command
```

S2.10 SpeciesRax and MiniNJ

SpeciesRax is part of GeneRax (version 2.0.1). We ran it with the following command:

```
mpiexec -n 16 generax --families famlies.txt --species-tree MiniNJ  \
    --strategy SKIP --rec-model UndatedDTL --skip-family-filtering \
    --do-not-reconcile  --prune-species-tree --per-family-rates \
    --prefix output-dir --si-strategy HYBRID
```

MiniNJ is run during SpeciesRax's execution in order to generate the starting species tree. To run MiniNJ by itself we replaced `--si-strategy HYBRID` with `--si-strategy SKIP`.

S3 Empirical Statistics of Datasets

Table S2: Empirical statistics for data simulated as part of the study. The information given in each row specifies how the model condition deviates from the default, which is 101 species, 1000 genes estimated from 100bp alignments, haploid effective population size of 5×10^7, duplication rate of 5×10^{-10}, and an equal rate of loss. The duplication rates are the following: low = 1×10^{-10}, mid = 5×10^{-10}, and high = 1×10^{-9}. The lowest and highest ILS conditions were generated using haploid population sizes of 1×10^4 and 2×10^8, respectively. MGTE refers to minimum gene tree estimation error, expressed in terms of the normalized Robinson-Foulds distance. The # leaves column reports the average number of leaves in each gene family tree across the replicates.

Dataset	# Species	# Genes	AD	MGTE	# leaves
50bp	101	1000	0.2030	0.5568	165.3
100bp (Default)	101	1000	0.2030	0.4336	165.3
500bp	101	1000	0.2030	0.1929	165.3
GDL (low dup rate; L/D =0)	101	1000	0.2397	0.4411	145.1
GDL (low dup rate; L/D = 0.5)	101	1000	0.2349	0.4494	128.0
GDL (low dup rate; L/D = 1)	101	1000	0.2341	0.4402	116.6
GDL (mid dup rate; L/D = 0)	101	1000	0.2659	0.4309	550.0
GDL (mid dup rate; L/D = 0.5)	101	1000	0.2271	0.4508	290.6
GDL (high dup rate; L/D = 0)	101	1000	0.2387	0.4718	3727.8
GDL (high dup rate; L/D = 0.5)	101	1000	0.2023	0.4339	993.0
GDL (high dup rate; L/D = 1)	101	1000	0.1906	0.3970	228.5
ILS (lowest ILS)	101	1000	0.0005	0.3775	164.9
ILS (highest ILS)	101	1000	0.5004	0.4392	170.1
10,000 Gene Trees (L/D = 0)	101	10,000	0.3002	0.4769	592.5
10,000 Gene Trees (L/D = 0.5)	101	10,000	0.2227	0.4380	311.2
10,000 Gene Trees (L/D = 1)	101	10,000	0.2124	0.4105	171.7
1000 Species	1001	1000	0.2349	0.4443	1578.1
Missing Data	101	1000	0.2030	0.4336	90.51

	0	0.5	1
1×10^{-10}	101	100	97
5×10^{-10}	101	96	72
1×10^{-9}	101	97	56

Table S3: Median number of species in the true gene family trees under different GDL model conditions (rows indicate gene duplication rates and columns indicate the relative probability of loss to duplication).

S4 Additional Results on Simulated Datasets

S4.1 Statistics about DISCO decomposition

In this section we explore properties about DISCO decompositions, including the size of DISCO trees. For CA-DISCO analyses, we include all the DISCO trees, but when we use DISCO with ASTRAL or ASTRID (i.e., with a summary method) we filter out the DISCO trees that have fewer than four species. For this reason, we obtain different average sizes for DISCO trees in the two settings: when used with concatenation analyses (CA-DISCO) and when used with summary methods.

Table S4: Empirical statistics of the CA-DISCO matrix of concatenated alignments (used in Experiment 2)—rows, columns, and percent gap characters—averaged over 10 replicates for each dataset. For context, the average number of leaves in the DISCO trees is also given. Datasets are identified by how they vary from the default conditions (101 species, 50 genes, 20% ILS, 100bp alignments, 5×10^{-10} Dup Rate, and an equal loss rate). Notably, trees with fewer than 4 taxa are *not filtered* from decomposition output when doing concatenation, as trees with less taxa could still contribute usable information.

Dataset	Rows	Columns	Gaps %	Avg. DISCO Tree Size
10 Genes	101	30,810	93.87 %	6.20
100 Genes	101	277,030	93.90 %	6.14
500 Genes	101	1,369,390	93.93 %	6.10
1000 Genes	101	2,709,750	93.93 %	6.10
50bp	101	80,365	94.53 %	5.48
100bp (Default)	101	144,720	93.95 %	6.08
500bp	101	629,450	93.06 %	6.99
Dup Rate 1×10^{-10}	101	26,080	78.75 %	21.20
Dup Rate 1×10^{-9}	101	276,560	95.98 %	4.02
ILS 0 %	101	116,930	93.42 %	6.64
ILS 50 %	101	141,670	93.90 %	6.07
10 Genes (Missing Data)	101	18,660	94.20 %	5.86
50 Genes (Missing Data)	101	79,340	94.37 %	5.95
100 Genes (Missing Data)	101	155,460	94.11 %	5.94
500 Genes (Missing Data)	101	776,330	94.15 %	5.90
1000 Genes (Missing Data)	101	1,539,220	94.12 %	5.92
1000 Species	1001	1,015,333	99.24 %	7.47

In Figure S1 we show the distribution of the number of species (same as number of leaves) in the largest (filtered)-DISCO tree and the number of species in the gene family trees (i.e., after filtering out the DISCO trees and gene family trees that have fewer than 4 species). We see that the number of species in the largest DISCO tree ranges substantially between model conditions, with larger numbers when losses do not occur (loss = 0 conditions), and then maximum sizes decreasing as the probability of losses increases. The number of species in each gene family tree also follows the same trends, with generally smaller gene family trees under high duplication rates with high loss rates. In Figure S2 we see that nearly all the species are contained in the largest tree outputted by DISCO for model conditions with low duplication rates and that at worse around 70% of the taxa are retained.

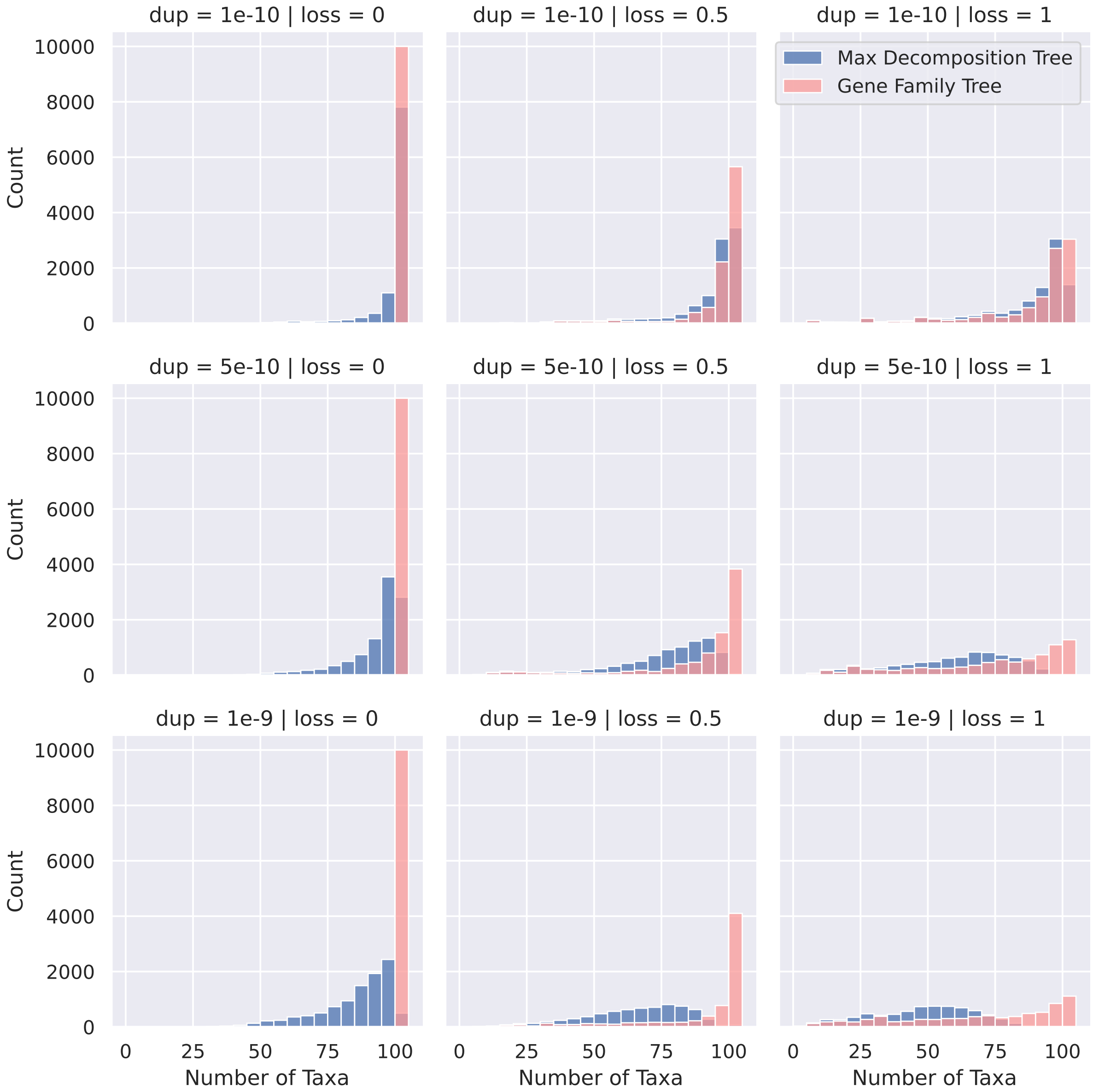


Figure S1: Histograms depicting the number of taxa (species) in the largest DISCO tree extracted from each gene family tree as well as the number of species in the gene-family trees across nine model conditions with varying duplication and loss rates. Results are gathered from 10 replicates per model condition with 101 species and 1000 genes each. Gene family trees and DISCO trees with fewer than 4 species are filtered.

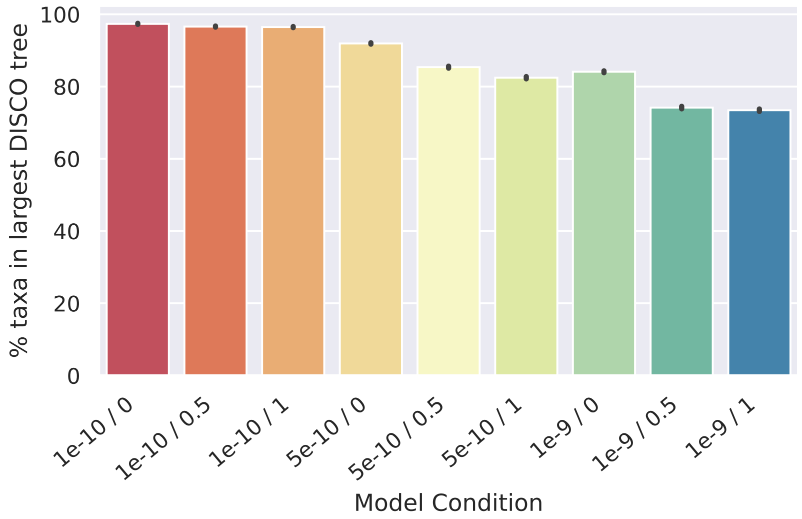


Figure S2: Percentage of species in the largest tree outputted by DISCO compared to the number of species in the gene family tree. Results are averaged over all gene family trees from all replicates (for a total of 10,000) for each model condition. The model conditions show three duplication rates and three loss rates (as dup / loss) for a total of nine conditions. Gene family trees and DISCO trees with fewer than 4 taxa are filtered. Note that the average percentage is close to 100% for the model conditions with low duplication rates and then drops to just above 70% for the condition with the highest duplication rate and equal loss rate.

	0	0.5	1
1×10^{-10}	2.6 / 52.8 / 145.1	2.5 / 50.1 / 130.6	2.4 / 47.4 / 120.9
5×10^{-10}	22.2 / 21.1 / 550.0	12.1 / 20.3 / 310.6	7.3 / 19.0 / 186.4
1×10^{-9}	207.8 / 12.8 / 3721.4	42.8 / 13.5 / 842.2	12.5 / 13.4 / 255.5

Table S5: Average number of (filtered)-DISCO trees produced for each gene family tree, the average number of leaves in those DISCO trees, and the average number of leaves in the original gene family tree under model conditions with 101 species, 20% AD, 43% MGTE, and different relative probabilities of loss (columns) to duplication (rows). Results shown here are given as $X/Y/Z$, where X denotes the number of DISCO trees, Y denotes the average number of leaves in those DISCO trees, and Z denotes the average number of leaves in the original gene family tree. DISCO trees and gene family trees with fewer than four species are filtered out.

S4.2 Peak memory usage

The peak memory results for some conditions were omitted from the paper as they continued to show the same trends already observed, and are provided here.

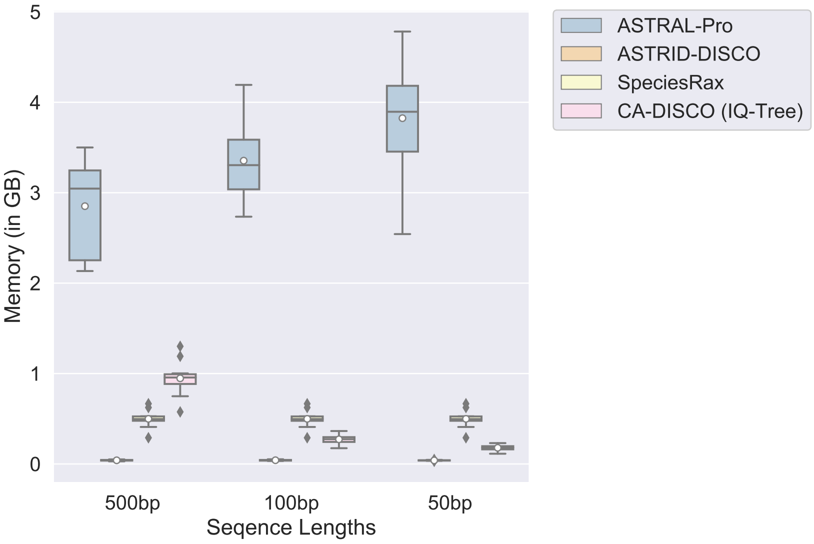


Figure S3: Peak memory used by methods (in GB) for differing levels of mean gene tree estimation error (MGTE); averages over 10 replicates per model condition are shown. All the datasets have 101 species, 50 gene trees, AD=20%, a duplication rate of 5.0×10^{-10} and an equal loss rate. The datasets include true gene trees and estimated gene trees from three sequence lengths (500bp, 100bp, and 50bp) and have MGTE of 19.2%, 43.3%, and 55.7% respectively. The boxes stretch from the 1st to 3rd quartile, the lines through the boxes show the medians, and the dots show means.

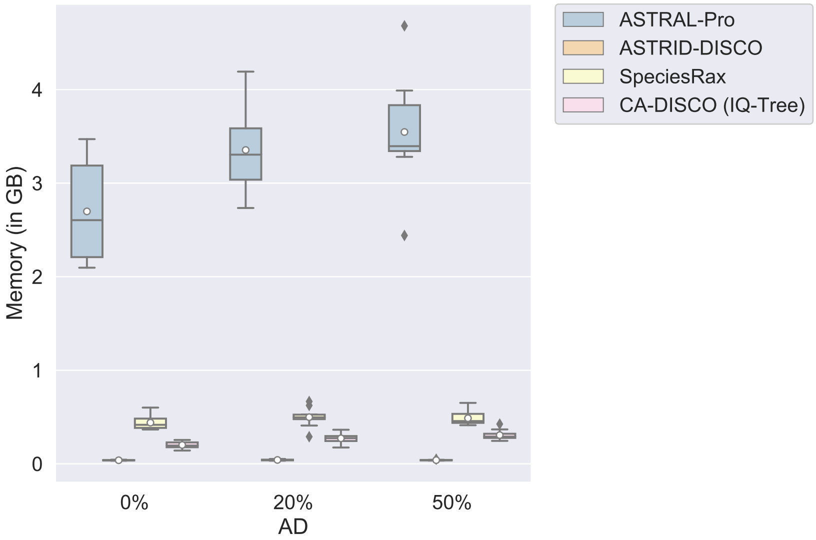


Figure S4: Peak memory used by methods (in GB) for differing levels of ILS; averages across 10 replicates per model condition are shown. All the datasets have 101 species, 50 gene trees estimated from 100bp alignments (approx. 43% MGTE), a duplication rate of 5.0×10^{-10}, and an equal loss rate.

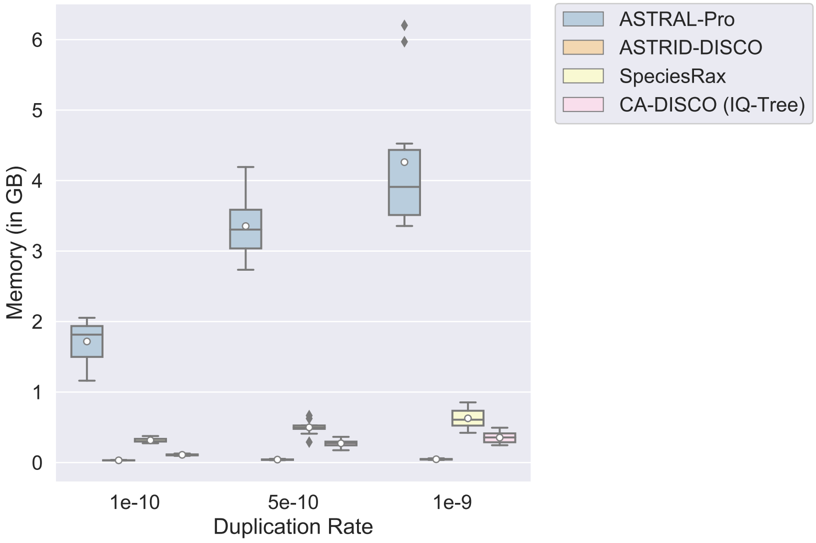


Figure S5: Peak memory used by methods (in GB) under differing duplication rates; averages across 10 replicates are shown. All the datasets have 101 species, 50 gene trees, gene trees estimated from 100bp alignments (43.3% MGTE), and AD=20%, and an equal loss rate to the duplication rate.

S4.3 Orthology estimation

The set of leaves in each DISCO tree can be considered a set of putative orthologs, and the set of DISCO trees produced by decomposing a gene family tree defines disjoint orthogroups. This estimation of orthology relationships can then be compared to the true pairwise orthology relationships (known when performing a simulation) and evaluated for accuracy. We can refer to the true relationships defined by the simulation as "actual orthologs" and "actual paralogs", and then use these to specify whether DISCO's orthology assessment is accurate (i.e., true positives and true negatives) or inaccurate (i.e., false positives and false negatives) for each pair of gene copies. The "true positives" are the pairs DISCO assesses to be orthologs which are actual orthologs, the "true negatives" are the pairs DISCO assesses to not be orthologs and which are actual paralogs, the "false positives" are the pairs DISCO assesses to be orthologs but which are actual paralogs, and the "false negatives" are the pairs DISCO assesses to not be orthologs but which are actually orthologs. However, since this decomposition depends on ASTRAL-Pro's rooting and tagging, to understand the orthology estimation accuracy obtained using DISCO, we consider the orthology accuracy for ASTRAL-Pro separately as well. We report both precision (i.e., TP/(TP+FP)) and recall (i.e., TP/(TP+FN)) for both ASTRAL-Pro and DISCO in Table S6. These results show that precision and recall are high but not perfect for ASTRAL-Pro on both true and estimated gene trees (with better results on the true gene trees, as expected). The imperfect results even on true gene trees is likely a combination of the existence of ILS and the technique ASTRAL-Pro uses to root the gene family trees. Precision and recall values for DISCO (based on the ASTRAL-Pro commands) show some differences to ASTRAL-Pro: the precision improves for DISCO compared to ASTRAL-Pro, but the recall drops; this is expected since DISCO can only recover some of the relationships because it partitions the leafset into disjoint sets.

Model Condition	ASTRAL-Pro	DISCO(ASTRAL-Pro)
Estimated Genes: Precision/Recall		
1.0×10^{-10} / 1	0.83 / 0.90	0.84 / 0.82
5.0×10^{-10} / 0	0.95 / 0.66	0.97 / 0.31
5.0×10^{-10} / 0.5	0.55 / 0.61	0.62 / 0.38
5.0×10^{-10} / 1	0.45 / 0.60	0.51 / 0.40
1.0×10^{-9} / 1	0.34 / 0.50	0.44 / 0.27
True Genes: Precision/Recall		
1.0×10^{-10} / 1	0.83 / 0.96	0.84 / 0.86
5.0×10^{-10} / 0	1.00 / 0.92	1.00 / 0.39
5.0×10^{-10} / 0.5	0.55 / 0.78	0.62 / 0.47
5.0×10^{-10} / 1	0.46 / 0.73	0.52 / 0.48
1.0×10^{-9} / 1	0.35 / 0.65	0.45 / 0.33

Table S6: Assessment of accuracy of ortholog detection. Precision / Recall rates are reported averaged over the total number of gene trees examined (10 replicates with 1000 genes each). Model conditions give the duplication rate / loss rate ratio. "ASTRAL-Pro" shows the accuracy of the orthologs detected by ASTRAL-Pro's rooting & tagging algorithm, while "DISCO(ASTRAL-Pro)" gives the accuracy of the orthologs retained after using DISCO to decompose the gene tree, using the ASTRAL-Pro tagging and rooting. The ILS level is moderate (AD=20%) and the estimated gene trees have 43% MGTE. All DISCO trees are considered irrespective of their size.

Method	Max Tree Size	Coverage	Orthology (Precision/Recall)
MI	17.3	0.62	0.76 / 0.32
DISCO	57.9	0.85	0.53 / 0.41

Table S7: Orthology prediction accuracy (precision and recall) and other statistics comparing MI and DISCO decompositions for the default model condition (10 replicates). We show the averages (across gene family trees) for maximum tree size (number of leaves) and coverage (proportion of gene family tree leafset contained in output produced by the decomposition method) in addition to the averages (across replicates) of precision/recall for orthology predictions produced by DISCO and MI. For max tree size and coverage, gene family trees with fewer than four species are excluded from the dataset; the average number of species in the gene family trees with at least four species is 73.1. Each replicate has 101 species, 100 genes, 20% AD, 43% MGTE, and a duplication rate of 5×10^{-10} with an equal loss rate.

S4.4 Missing Data

In our paper, we discussed the cause for missing data, focusing on the CA-DISCO matrix. Here we provide additional discussion and results regarding this question and related questions.

In our study, close to 100% of the replicates had all the species in at least one of the DISCO decomposition subtrees; the only cases where this was not true were ones with only ten genes, where (for a few replicates) we produced outputs with 100 instead of 101 species. Thus, the DISCO decomposition only rarely produces an output that completely omits some species.

The clade-based missing data condition explores the impact of removing species from a given gene that are not contained within a randomly selected clade in the species tree. This is equivalent to exploring the impact of gene birth below the root of the species tree (i.e., "late gene birth"). This experiment was restricted to the default model condition, and so we report results only for this case. Late gene birth reduces the number of species fairly substantially: before the missing data process is applied, we had a median (across the genes) of 72 species for the default model condition, and afterwards the median (among impacted gene families) was only 28 (on average 41.8% of species were removed from affected gene families). This is a large impact, but as noted in the main paper this process affects only a fraction of gene families (60.2%), and so the overall impact is small.

To assess the impact of GDL on missing data, we examine the number of species in each of the true gene family trees. These range in terms of the number of species, as shown in Table S3. As expected, when there is no loss, then all gene family trees have all the species (101). However, as the loss rate increases, the median number of species per gene family tree decreases, so that for our default model condition the median number of species per gene family tree is only 72.

These analyses show that the GDL process can have a large impact on individual gene family trees (but this depends on the GDL model, with a larger impact when loss rates equal duplication rates combined with high rates of duplication). However, when considered across all the genes, there is almost always some gene containing a given species, and so the overall impact is small.

Late gene birth similarly has a large impact on individual genes impacted by the late birth, but overall has a small impact on the full set of genes (since most genes are not impacted by late birth). In contrast, the DISCO decomposition has a very large impact on all gene family trees we examined (the exception would be cases where there is no GDL), and this impact is larger than for both GDL and late gene birth.

S5 Differences between ASTRAL-Pro and ASTRAL-DISCO

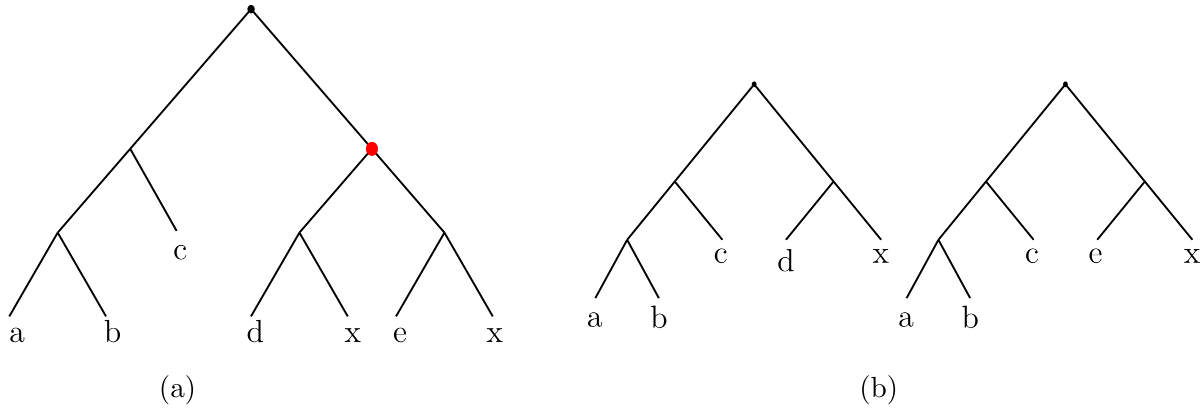


Figure S6: An input where ASTRAL-Pro scores more quartets than ASTRAL-DISCO. *(a)* Given an unrooted tree, ASTRAL-Pro will root and tag it to minimize the total number of gene duplications and losses. The tree shown here is one of the possible optimal outcomes of this rooting-and-tagging procedure when ASTRAL-Pro is provided with the unrooted version of the tree (the red dots indicate duplication nodes). *(b)* The two possible DISCO decompositions of the gene tree from (a) (drawn as rooted trees); note that $ab|de$ does not appear in either of these trees, and hence $ab|de$ will not be used by ASTRAL-DISCO in computing a species tree. Similarly, neither of the decompositions that is produced contains both d and e; hence, exactly one of $ab|cd$ and $ab|ce$ will be scored by ASTRAL-DISCO. However, these quartet trees ($ab|de, ab|cd, ab|ce$) are each valid speciation-driven quartets, and so will be scored by ASTRAL-Pro.

Every quartet scored by ASTRAL-DISCO is also scored by ASTRAL-Pro, but the converse does not always hold; hence, ASTRAL-Pro has the potential to use more of the information in the input gene family trees. Here we explain why this can happen, and we provide an example of such a case.

Given a rooted and tagged gene tree, ASTRAL-Pro scores a quartet tree q if this quartet is speciation-driven (meaning that it has four distinct species labels and the least common ancestor (LCA) of any three leaves is a speciation node). Since DISCO uses the rooting and tagging provided by ASTRAL-Pro, this implies that ASTRAL-DISCO considers a subset of the quartet trees considered by ASTRAL-Pro. In Figure S6, we show an example of an input gene family tree on six species with one duplication event where ASTRAL-DISCO will only use a proper subset of the quartet trees considered by ASTRAL-Pro.

S6 Results on Biological Datasets

Each estimated phylogeny is rooted at the established outgroup. We compare these rooted topologies and report differences between the ASTRAL-Pro, SpeciesRax, and Astrid-DISCO trees. We also discuss results more generally with reference to current taxonomic knowledge. We computed the clade difference for each pair of the rooted trees using the `comparePhylo` function from the R package APE [11].

These analyses show that the conflicts between the trees correspond to relationships that have long been discussed in the literature. Furthermore, the branch support on the estimated trees (computed using ASTRAL's posterior probability branch support technique (i.e., localPP) [12]) are nearly all very high. Overall, all three methods produced very high support edges across all biological datasets. For the Fungi dataset, all edges have 100% localPP support. In the 1KP and Vertebrate188 datasets, edges presenting less than 0.99 support are all conflicting edges between methods and the literature, as we discuss below. The sole exception is the Zygnematophyceae clade (in the 1KP dataset) for which all methods returned the same topology but with some nodes having 100% localPP support.

S6.1 Fungi Dataset

As seen in Figure S7, all the methods produce the same tree topology (which is also are identical to the FastMulRFS tree reported in [6]), and differ from the reference tree reported in [13] on two edges. All the edges in this common tree have posterior probability of 100%.

S6.2 1KP Plant Dataset

All tested methods returned a different species tree (Figs. S8 - S10). We rooted all trees on the Chlorophyta clade (*Chlamydomonas* + *Volvox* + *Uronema*), as in [14].

- *Early diverging Streptophyta.* A-Pro returned a tree where Charales (only represented by *Chara* in this dataset) are diverging after Coleochaetales. Although Charales have initially been thought to be the sister-group of the land plants (see a review in [15]), it now seems widely accepted that Coleochaetales are closer to land plants than Charales are [16–19], thus favoring the SpeciesRax and ASTRID-DISCO topologies with respect to that conflict.

- *Three Major Angiosperm Groups.* While ASTRAL-Pro and SpeciesRax recovered [monocots [Eudicots + magnoliids]], ASTRID-DISCO recovered [Eudicots [monocots + magnoliids]] tree. The relationships between Eudicots, Magnoliids and Monocots have been subject to debate since the beginning of Angiosperm classification. The presence of *Ceratophyllum* and Chloranthaceae seems to be essential to assert the relationships between those three major groups, and this dataset only comprises one specimen of Chloranthaceae, making it hard to trust either of the two configurations.

- *Position of Vitis* SpeciesRax is the only method that correctly placed *Vitis* among Rosiids, while ASTRAL-Pro and ASTRID-DISCO recovered *Vitis* as sister-group to core-eudicots, as in [14]. We note that [14] pointed out that the placement of *Vitis* can be problematic when taxon sampling is poor (see [20]).

- *Monocot relationships.* ASTRAL-Pro and ASTRID-DISCO both support the [commelinids [Asparagales + Liliales]] hypothesis while SpeciesRax supports the [Liliales [Asparagales + commelinids]] hypothesis. The [Liliales [Asparagales + commelinids]] hypothesis is in agreement with the current consensus [21] but the recent study of [18] found support for the [commelinids [Asparagales + Liliales]] hypothesis.

- *Position of Zygophyllales.* Here Zygophyllales is only represented by *Larrea*. Currently considered as part of the Fabids, *Larrea* was nevertheless found to be sister to Malvids in both the ASTRAL-Pro and ASTRID-DISCO analyses, and only found sister to the rest of Fabids in the SpeciesRax tree. In the most up-to-date phylogeny of Rosiids [22], it is not clear whether the Zygophyllales are sister to Malvids or within Fabids, and this seems to be dependent on the type of data used.

- *Lamiids relationships.* Both ASTRAL-Pro and SpeciesRax found the [Apocynaceae[*Rosmarinus* +*Ipomoea*]] hypothesis while ASTRID-DISCO recovered a less supported [*Rosmarinus*[Apocynaceae+*Ipomoea*]] hypothesis. The most up-to-date phylogeny [23] found a highly supported [*Rosmarinus*[Apocynaceae+*Ipomoea*]] relationship, thus supporting the ASTRID-DISCO results, even though it is less supported.

S6.3 Vertebrate188

All tested methods returned a different species tree (Figs. S11–S13). We rooted all trees on the jawless fishes node (*Eptatretus* + *Petromyzon*), the sister group of all jawed vertebrates.

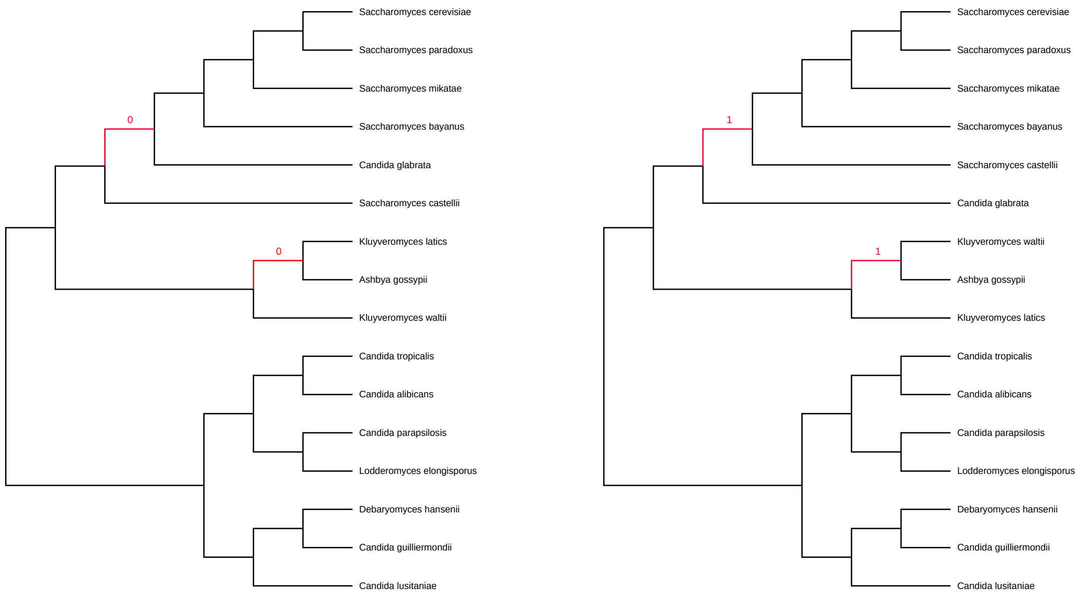
- *Relationships between Eupercaria, Anabantaria, Carangaria and Ovalentaria, sensu [24].* Both SpeciesRax and ASTRID-DISCO retrieved a topology where Ovalentaria are sister to (Eupercaria+ Anabantaria/Carangaria) while ASTRAL-Pro found Eupercaria to be sister to (Ovalentaria+Anabantaria/Carangaria). The latter hypothesis is in agreement with the current phylogenetic classification [24] as well as with more recent studies [25, 26].

- *Position of clingfishes (represented here by Gouania).* ASTRAL-Pro and ASTRID-DISCO analyses recovered *Gouania* sister to all other Ovalentaria representatives, while SpeciesRax recovered it within Ovalentaria, as sister to the Beloniformes/Cyprinodontiformes clade. As we are missing several representatives of this hyperdiverse group of fishes and as the basal relationships at the Ovalentaria node are often not supported (e.g. [24]), even with phylogenomic data (e.g. [27]), it is currently impossible to favor one (if any) of the two hypothesis.

- *Position of Scadentia, Chiroptera and Perissodactyla.* These three groups have in common a longstanding debate on their phylogenetic position within Euarchontoglires (for Scadentia) or within Laurasiatheria (for Chiroptera and Perissodactyla), as discussed in a recent review by [28]. Despite extensive phylogenomic studies with sometimes thousands of loci (e.g. [29, 30]), there is still no consensus on the placement of these taxa.

- *Conflicts between Sciuromorpha, Hystricomorpha and Castorimorpha/Myomorpha.* The relationships between those major Rodent lineages have also been subject to a long debate. However, some of the latest phylogenomic studies seem to agree with a [Castorimorpha/Myomorpha [Hystricomorpha + Sciuromorpha]] hypothesis [30–32], thus favoring the SpeciesRax topology even though it is the less supported one.

- *New World Monkeys relationships.* Three recent studies [33–35] have investigated the relationships between the three major Cebidae lineages – Cebinae, Aotinae and Callitrichinae. Unsurprisingly, there is still no consensus for the relationships between those three lineages.



(a) Reference Tree

(b) ASTRAL-Pro / ASTRID-DISCO / SpeciesRax

Figure S7: *Experiment 3.* 16-taxon fungi dataset from [2]. All methods produce a tree containing two branches which disagree with the reference from [13]. Disagreeing branches are highlighted in red. Local posterior probability (localPP) branch support [12] computed by ASTRAL-Pro is shown for the highlighted branches. (All branches in the estimated tree have localPP support of 1.0.)

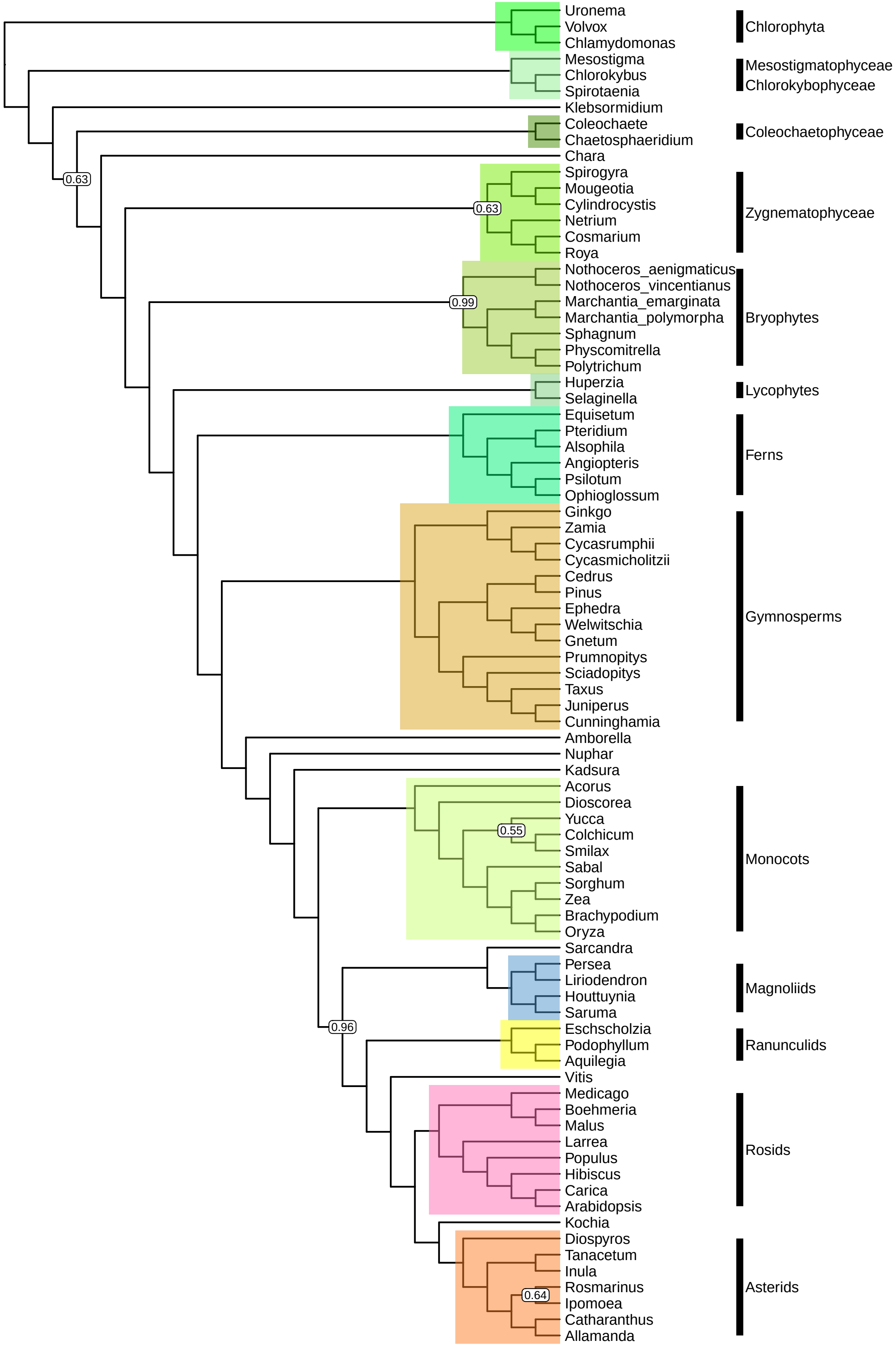


Figure S8: Species tree returned by ASTRAL-Pro on the 83-taxon plant dataset (1KP). All branches in the estimated tree have localPP support of 1.0 except for those branches that are labeled with numbers.

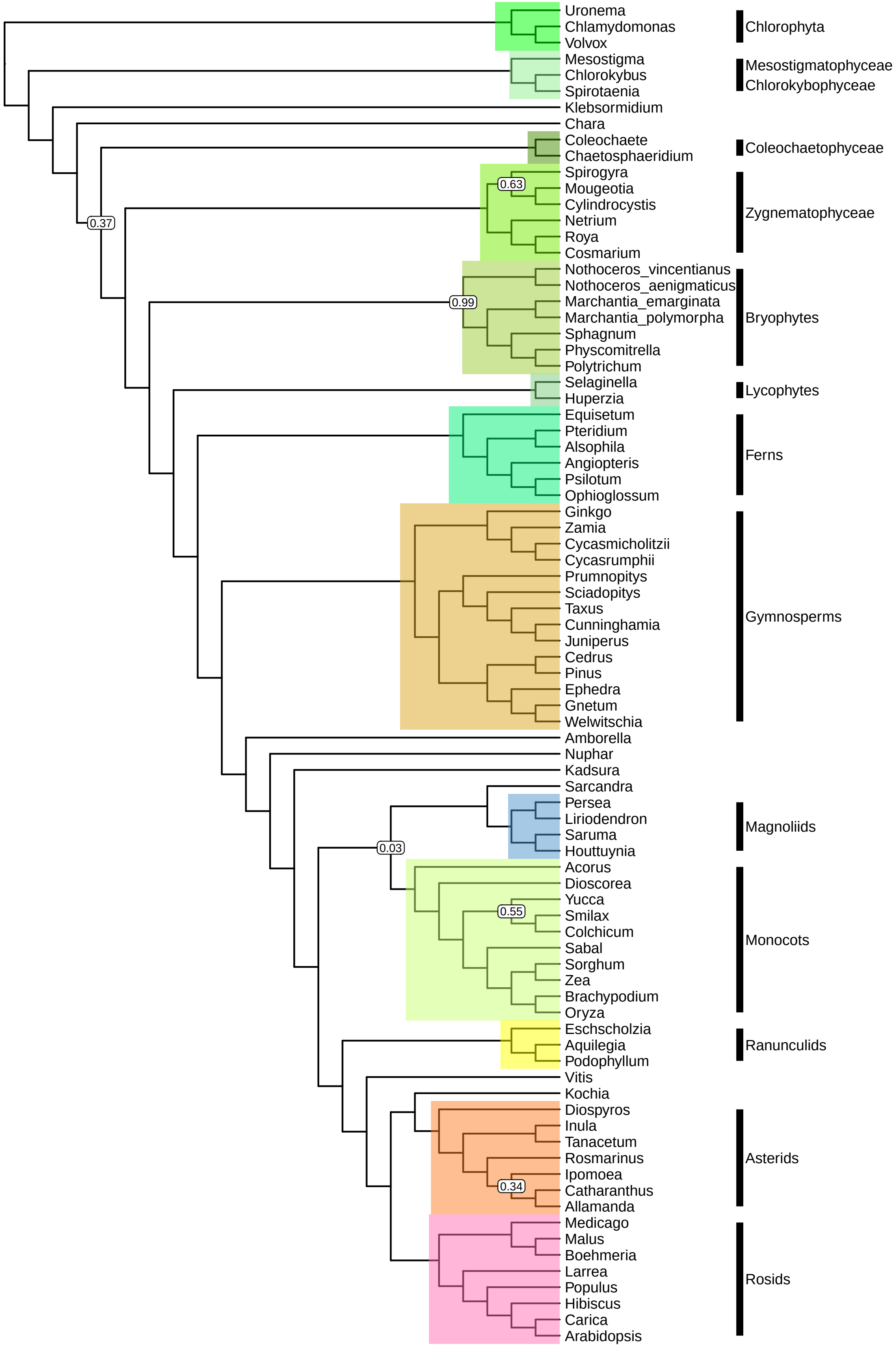


Figure S9: Species tree returned by ASTRID-DISCO on the 83-taxon plant dataset (1KP). All branches in the estimated tree have localPP support of 1.0 except for those branches that are labeled with numbers.

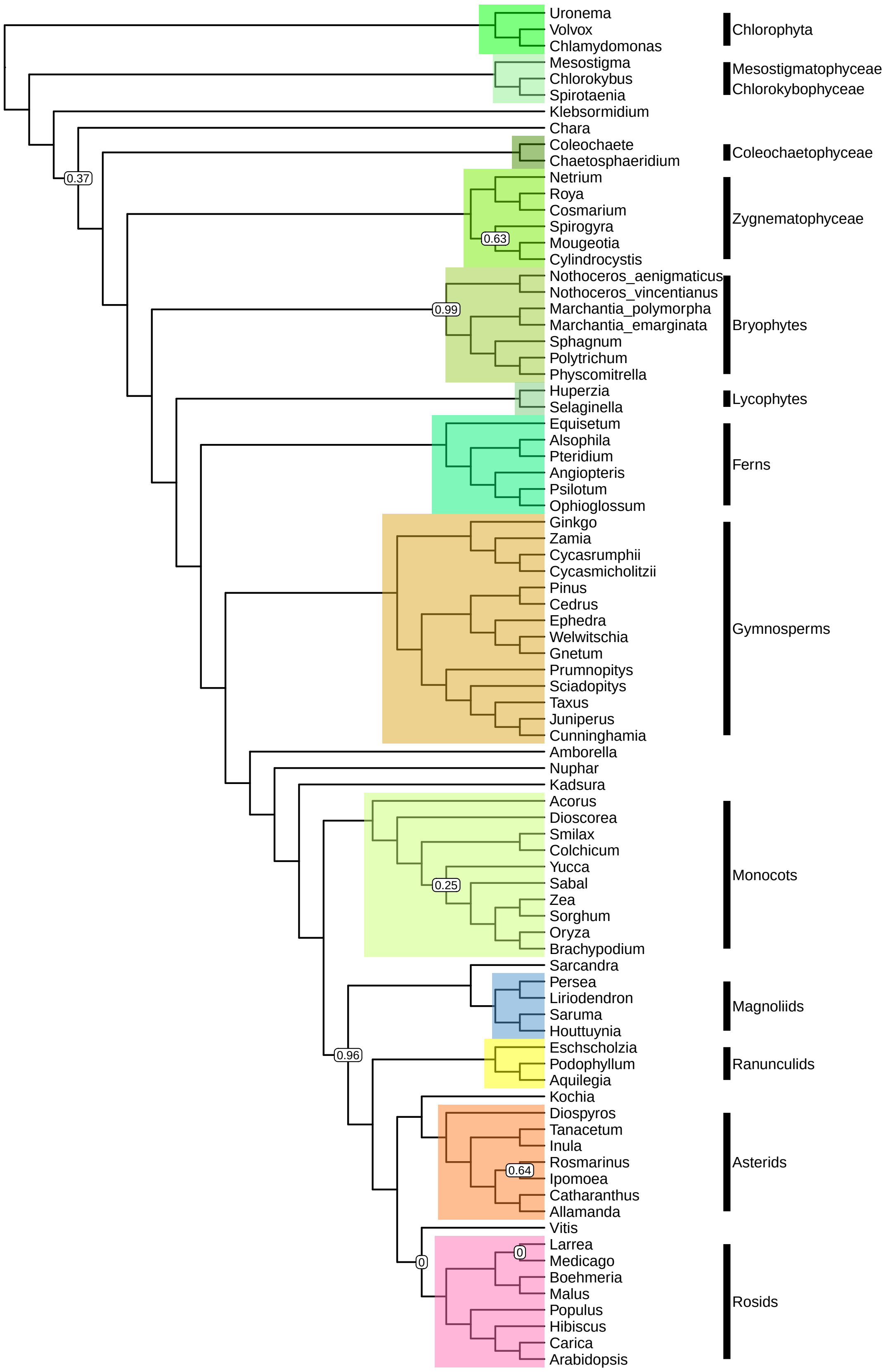


Figure S10: Species tree returned by SpeciesRax on the 83-taxon plant dataset (1KP). All branches in the estimated tree have localPP support of 1.0 except for those branches that are labeled with numbers.

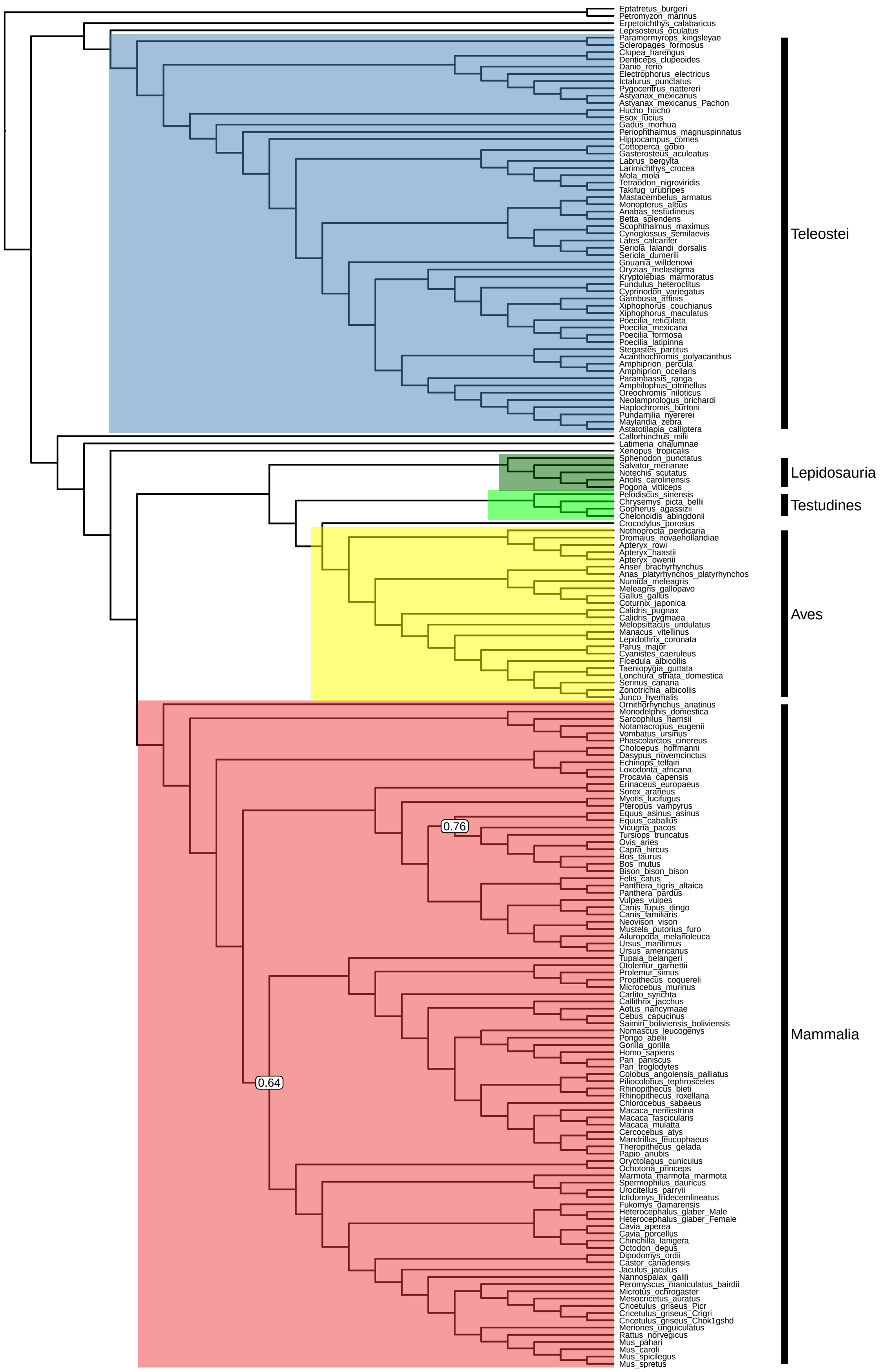


Figure S11: Species tree returned by ASTRAL-Pro on the 188-taxon vertebrate dataset. All branches in the estimated tree have localPP support of 1.0 except for those branches that are labeled with numbers.

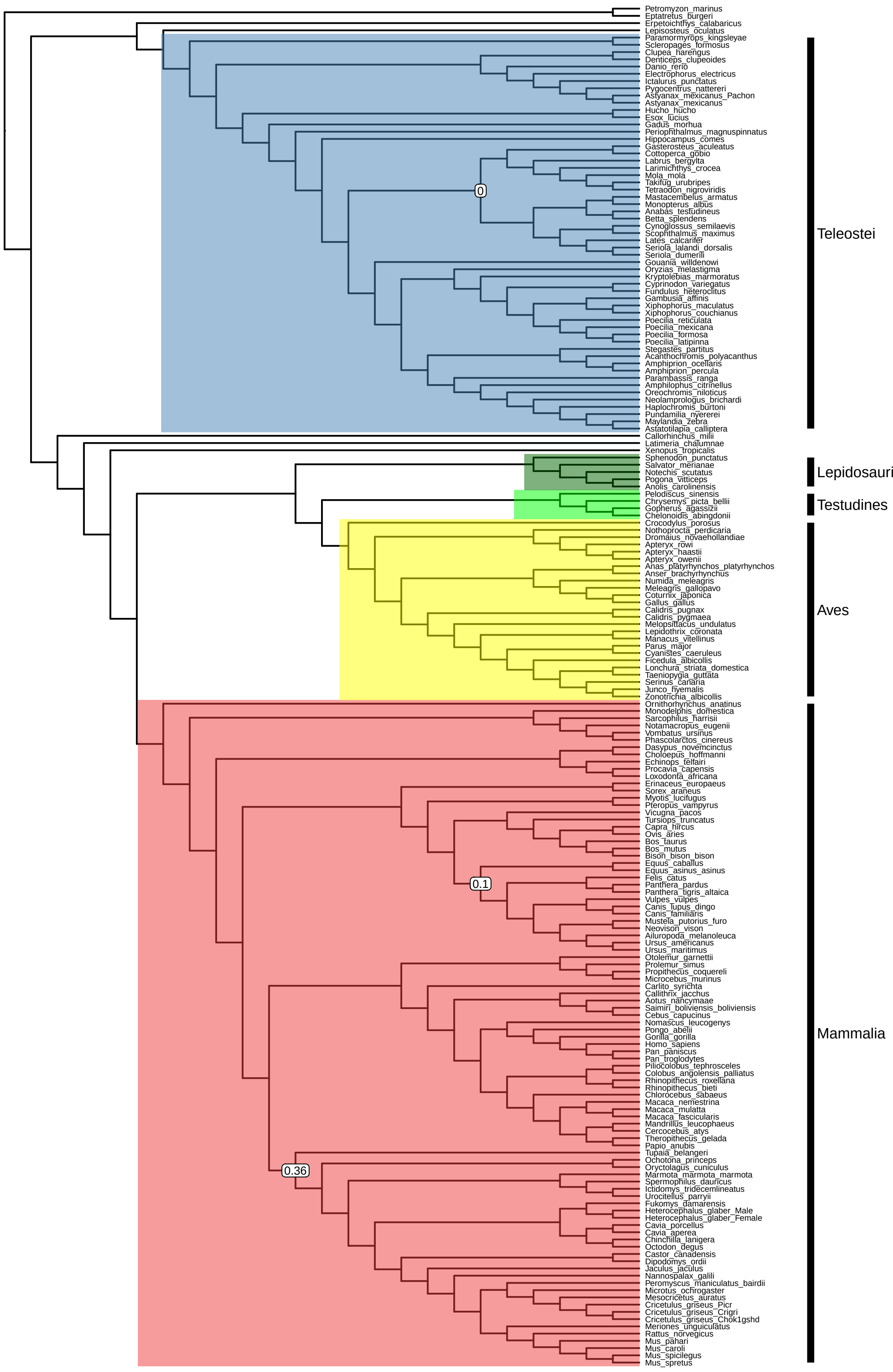


Figure S12: Species tree returned by ASTRID-DISCO on the 188-taxon vertebrate dataset. All branches in the estimated tree have localPP support of 1.0 except for those branches that are labeled with numbers.

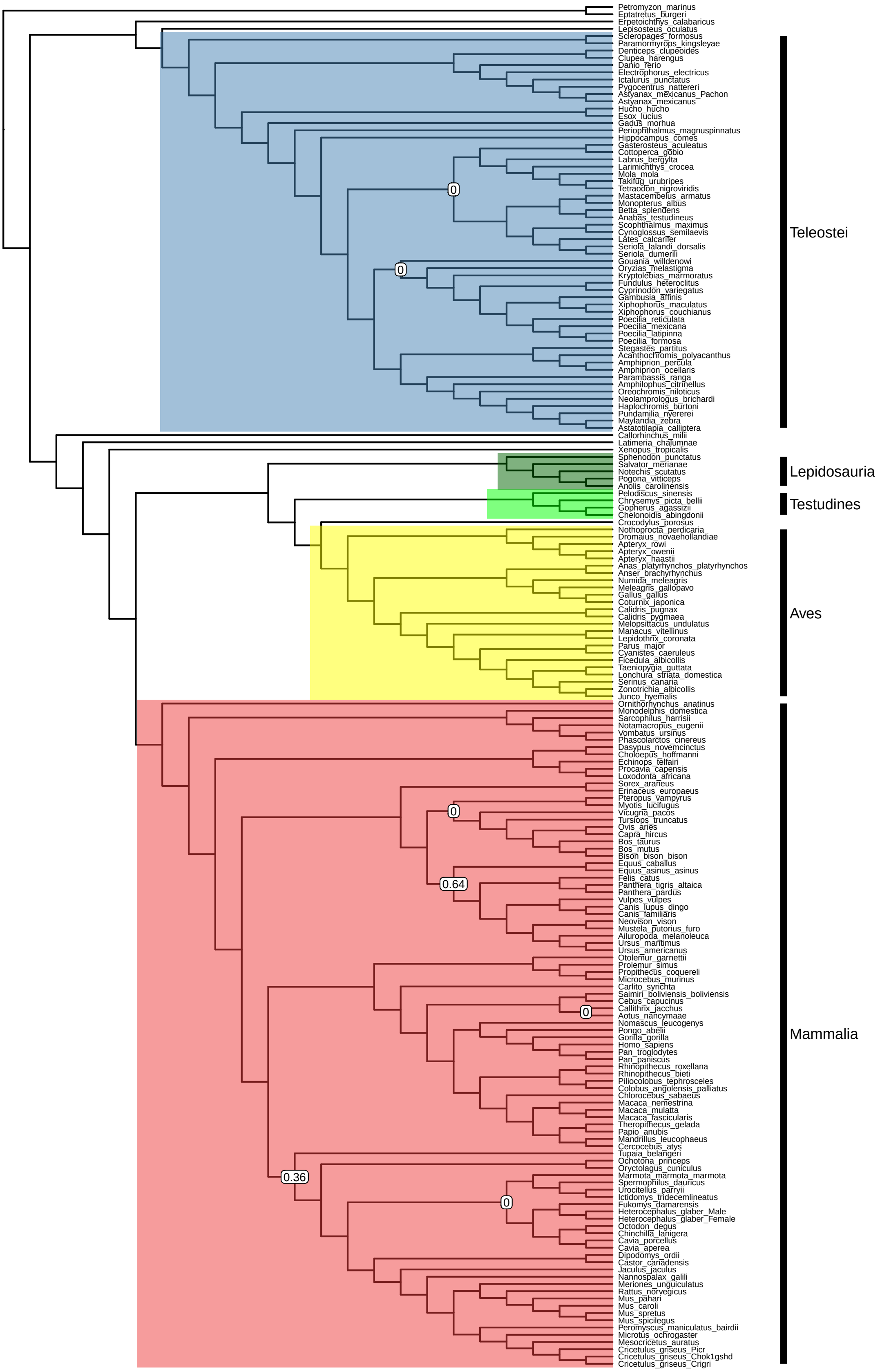


Figure S13: Species tree returned by SpeciesRax on the 188-taxon vertebrate dataset. All branches in the estimated tree have localPP support of 1.0 except for those branches that are labeled with numbers.

Conflicts	Method's Result	Support
ASTRAL-Pro		
Divergence of Chara and Coleochaetales	Coleochaetales-first	0.63
3 Major Angiosperm Groups	[monocots [Eudicots + magnolids]]	0.96
Position of Vitis	Sister to core-eudicots	1
Monocot relationships	[commelinids [Asparagales + Liliales]]	0.55
Lamiids relationships.	[Apocynaceae[Rosmarinus +Ipomoea]]	0.64
Position of Zygophyllales	sister to Malvids	1
SpeciesRax		
Divergence of Chara and Coleochaetales	Chara-first	0.37
3 Major Angiosperm Groups	[monocots [Eudicots + magnolids]]	0.96
Position of Vitis	early diverging rosiids	0
Monocot relationships	[Liliales [Asparagales + commelinids]]	0.25
Lamiids relationships.	[Apocynaceae[Rosmarinus +Ipomoea]]	0.64
Position of Zygophyllales	sister to Fabids	0
ASTRID-DISCO		
Divergence of Chara and Coleochaetales	Chara-first	0.37
3 Major Angiosperm Groups	[Eudicots [monocots + magnolids]]	0.03
Position of Vitis	Sister to core-eudicots	1
Monocot relationships	[commelinids [Asparagales + Liliales]]	0.55
Lamiids relationships.	[Rosmarinus[Apocynaceae+Ipomoea]]	0.34
Position of Zygophyllales	Sister to Malvids	1

Table S8: Description of differences in species trees outputted by ASTRAL-Pro, SpeciesRax, and ASTRID-DISCO, on the 1KP dataset. Branch support is calculated using localPP in ASTRAL.

Conflicts	Method's Result	Support
	ASTRAL-Pro	
Relationship between Eupercaria, Anabantaria, Carangaria, and Ovalentaria	Eupercaria sister to the rest	1
position of Gouania	sister to rest of Ovalentaria	1
position of Tupaia	sister to Primates	0.64
position of Chiroptera	sister to Ferungulata	1
position of Perissodactyla	sister to Artiodactyla	0.76
Rodentia relationships	[Sciuromorpha [Hystricomorpha + Castorimorpha/Myomorpha]]	1
New World Monkeys relationships	[Callithrix [Aotus + Cebidae]]	1
	SpeciesRax	
Relationship between Eupercaria, Anabantaria, Carangaria, and Ovalentaria	Ovulentaria sister to the rest	0
position of Gouania	sister to Atherinomorphae	0
position of Tupaia	sister to Rodents	0.36
position of Chiroptera	sister to Artiodactyla	0
position of Perissodactyla	sister to Carnivora	0.64
Rodentia relationships	[Castorimorpha/Myomorpha [Hystricomorpha + Sciuromorpha]]	0
New World Monkeys relationships	[Cebidae [Callithrix + Aotus]]	0
	ASTRID-DISCO	
Relationship between Eupercaria, Anabantaria, Carangaria, and Ovalentaria	Ovulentaria sister to the rest	0
position of Gouania	sister to rest of Ovalentaria	1
position of Tupaia	sister to Rodents	0.36
position of Chiroptera	sister to Ferungulata	1
position of Perissodactyla	sister to Carnivora	0.11
Rodentia relationships	[Sciuromorpha [Hystricomorpha + Castorimorpha/Myomorpha]]	1
New World Monkeys relationships	[Callithrix [Aotus + Cebidae]]	1

Table S9: Description of differences in species trees outputted by ASTRAL-Pro, SpeciesRax, and ASTRID-DISCO, on the 188-taxon vertebrates dataset. Branch support is calculated using localPP in ASTRAL.

References

[1] Diego Mallo, Leonardo de Oliveira Martins, and David Posada. SimPhy: phylogenomic simulation of gene, locus, and species trees. *Systematic Biology*, 65(2):334–344, 2016.

[2] Matthew D Rasmussen and Manolis Kellis. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research*, 22(4):755–765, 2012.

[3] Chao Zhang, Celine Scornavacca, Erin K Molloy, and Siavash Mirarab. ASTRAL-Pro: Quartet-based species-tree inference despite paralogy. *Molecular Biology and Evolution*, 37(11):3292–3307, 2020.

[4] William Fletcher and Ziheng Yang. INDELible: a flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, 26(8):1879–1888, 2009.

[5] Simon Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, 17(2):57–86, 1986.

[6] Erin K Molloy and Tandy Warnow. FastMulRFS: fast and accurate species tree estimation under generic gene duplication and loss models. *Bioinformatics*, 36(Supplement_1):i57–i65, 2020.

[7] Pranjal Vachaspati and Tandy Warnow. FastRFS: fast and accurate Robinson-Foulds Supertrees using constrained exact optimization. *Bioinformatics*, 33(5):631–639, 2017.

[8] Siavash Mirarab, Rezwana Reaz, Md S Bayzid, Théo Zimmermann, M Shel Swenson, and Tandy Warnow. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, 2014.

[9] Pranjal Vachaspati and Tandy Warnow. SIESTA: enhancing searches for optimal supertrees and species trees. *BMC Genomics*, 19(5):252, 2018.

[10] Ya Yang and Stephen A Smith. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution*, 31(11):3081–3092, 2014.

[11] Emmanuel Paradis and Klaus Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3):526–528, 2019.

[12] Erfan Sayyari and Siavash Mirarab. Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution*, 33(7):1654–1668, 2016.

[13] Geraldine Butler, Matthew D Rasmussen, Michael F Lin, Manuel AS Santos, Sharadha Sakthikumar, Carol A Munro, Esther Rheinbay, Manfred Grabherr, Anja Forche, Jennifer L Reedy, et al. Evolution of pathogenicity and sexual reproduction in eight candida genomes. *Nature*, 459(7247):657–662, 2009.

[14] Norman J Wickett, Siavash Mirarab, Nam Nguyen, Tandy Warnow, Eric Carpenter, Naim Matasci, Saravanaraj Ayyampalayam, Michael S Barker, J Gordon Burleigh, Matthew A Gitzendanner, et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111(45):E4859–E4868, 2014.

[15] Bojian Zhong, Linhua Sun, and David Penny. The origin of land plants: a phylogenomic perspective. *Evolutionary bioinformatics*, 11:EBO–S29089, 2015.

[16] Jennifer L Morris, Mark N Puttick, James W Clark, Dianne Edwards, Paul Kenrick, Silvia Pressel, Charles H Wellman, Ziheng Yang, Harald Schneider, and Philip CJ Donoghue. The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences*, 115(10):E2274–E2283, 2018.

[17] Mark N Puttick, Jennifer L Morris, Tom A Williams, Cymon J Cox, Dianne Edwards, Paul Kenrick, Silvia Pressel, Charles H Wellman, Harald Schneider, Davide Pisani, et al. The interrelationships of land plants and the nature of the ancestral embryophyte. *Current Biology*, 28(5):733–745, 2018.

[18] James H. Leebens-Mack, Michael S. Barker, Eric J. Carpenter, Michael K. Deyholos, Matthew A. Gitzendanner, Sean W. Graham, Ivo Grosse, Zheng Li, Michael Melkonian, Siavash Mirarab, Martin Porsch, Marcel Quint, Stefan A. Rensing, Douglas E. Soltis, Pamela S. Soltis, Dennis W. Stevenson, Kristian K. Ullrich, Norman J. Wickett, Lisa DeGironimo, Patrick P. Edger, Ingrid E. Jordon-Thaden, Steve Joya, Tao Liu, Barbara Melkonian, Nicholas W. Miles, Lisa Pokorny, Charlotte Quigley, Philip Thomas, Juan Carlos Villarreal, Megan M. Augustin, Matthew D. Barrett, Regina S. Baucom, David J. Beerling, Ruben Maximilian Benstein, Ed Biffin, Samuel F. Brockington, Dylan O. Burge, Jason N. Burris, Kellie P. Burris, Valérie Burtet-Sarramegna, Ana L. Caicedo, Steven B. Cannon, Zehra Çebi, Ying Chang, Caspar Chater, John M. Cheeseman, Tao Chen, Neil D. Clarke, Harmony Clayton, Sarah Covshoff, Barbara J. Crandall-Stotler, Hugh Cross, Claude W. dePamphilis, Joshua P. Der, Ron Determann, Rowan C. Dickson, Verónica S. Di Stilio, Shona Ellis, Eva Fast, Nicole Feja, Katie J. Field, Dmitry A. Filatov, Patrick M. Finnegan, Sandra K. Floyd, Bruno Fogliani, Nicolás García, Gildas Gâteblé, Grant T. Godden, Falicia (Qi Yun) Goh, Stephan Greiner, Alex Harkess, James Mike Heaney, Katherine E. Helliwell, Karolina Heyduk, Julian M. Hibberd, Richard G. J. Hodel, Peter M. Hollingsworth, Marc T. J. Johnson, Ricarda Jost, Blake Joyce, Maxim V. Kapralov, Elena Kazamia, Elizabeth A. Kellogg, Marcus A. Koch, Matt Von Konrat, Kálmán Könyves, Toni M. Kutchan, Vivienne Lam, Anders Larsson, Andrew R. Leitch, Roswitha Lentz, Fay-Wei Li, Andrew J. Lowe, Martha Ludwig, Paul S. Manos, Evgeny Mavrodiev, Melissa K. McCormick, Michael McKain, Tracy McLellan, Joel R. McNeal, Richard E. Miller, Matthew N. Nelson, Yanhui Peng, Paula Ralph, Daniel Real, Chance W. Riggins, Markus Ruhsam, Rowan F. Sage, Ann K. Sakai, Moira Scascitella, Edward E. Schilling, Eva-Marie Schlösser, Heike Sederoff, Stein Servick, Emily B. Sessa, A. Jonathan Shaw, Shane W. Shaw, Erin M. Sigel, Cynthia Skema, Alison G. Smith, Ann Smithson, C. Neal Stewart, John R. Stinchcombe, Peter Szövényi, Jennifer A. Tate, Helga Tiebel, Dorset Trapnell, Matthieu Villegente, Chun-Neng Wang, Stephen G. Weller, Michael Wenzel, Stina Weststrand, James H. Westwood, Dennis F. Whigham, Shuangxiu Wu, Adrien S. Wulff, Yu Yang, Dan Zhu, Cuili Zhuang, Jennifer Zuidof, Mark W. Chase, J. Chris Pires, Carl J.

Rothfels, Jun Yu, Cui Chen, Li Chen, Shifeng Cheng, Juanjuan Li, Ran Li, Xia Li, Haorong Lu, Yanxiang Ou, Xiao Sun, Xuemei Tan, Jingbo Tang, Zhijian Tian, Feng Wang, Jun Wang, Xiaofeng Wei, Xun Xu, Zhixiang Yan, Fan Yang, Xiaoni Zhong, Feiyu Zhou, Ying Zhu, Yong Zhang, Saravanaraj Ayyampalayam, Todd J. Barkman, Nam-phuong Nguyen, Naim Matasci, David R. Nelson, Erfan Sayyari, Eric K. Wafula, Ramona L. Walls, Tandy Warnow, Hong An, Nils Arrigo, Anthony E. Baniaga, Sally Galuska, Stacy A. Jorgensen, Thomas I. Kidder, Hanghui Kong, Patricia Lu-Irving, Hannah E. Marx, Xinshuai Qi, Chris R. Reardon, Brittany L. Sutherland, George P. Tiley, Shana R. Welles, Rongpei Yu, Shing Zhan, Lydia Gramzow, Günter Theißen, Gane Ka-Shu Wong, and One Thousand Plant Transcriptomes Initiatives. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574(7780):679–685, 2019.

[19] Danyan Su, Lingxiao Yang, Xuan Shi, Xiaoya Ma, Xiaofan Zhou, S Blair Hedges, and Bojian Zhong. Large-scale phylogenomic analyses reveal the monophyly of bryophytes and Neoproterozoic origin of land plants. *Molecular Biology and Evolution*, 2021.

[20] Robert K Jansen, Charalambos Kaittanis, Christopher Saski, Seung-Bum Lee, Jeffrey Tomkins, Andrew J Alverson, and Henry Daniell. Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evolutionary Biology*, 6(1):1–14, 2006.

[21] Mark W Chase, MJM Christenhusz, MF Fay, JW Byng, Walter S Judd, DE Soltis, DJ Mabberley, AN Sennikov, Pamela S Soltis, and Peter F Stevens. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, 181(1):1–20, 2016.

[22] Miao Sun, Ryan A Folk, Matthew A Gitzendanner, Stephen A Smith, Charlotte Germain-Aubrey, Robert P Guralnick, Pamela S Soltis, Douglas E Soltis, and Zhiduan Chen. Exploring the phylogeny of rosids with a five-locus supermatrix from GenBank. *bioRxiv*, page 694950, 2019.

[23] Caifei Zhang, Taikui Zhang, Federico Luebert, Yezi Xiang, Chien-Hsun Huang, Yi Hu, Mathew Rees, Michael W Frohlich, Ji Qi, Maximilian Weigend, et al. Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. *Molecular biology and evolution*, 37(11):3188–3210, 2020.

[24] Ricardo Betancur-R, Edward O Wiley, Gloria Arratia, Arturo Acero, Nicolas Bailly, Masaki Miya, Guillaume Lecointre, and Guillermo Orti. Phylogenetic classification of bony fishes. *BMC evolutionary biology*, 17(1):1–40, 2017.

[25] Michael E Alfaro, Brant C Faircloth, Richard C Harrington, Laurie Sorenson, Matt Friedman, Christine E Thacker, Carl H Oliveros, David Černỳ, and Thomas J Near. Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary. *Nature Ecology & Evolution*, 2(4):688–696, 2018.

[26] Lily C Hughes, Guillermo Ortí, Hadeel Saad, Chenhong Li, William T White, Carole C Baldwin, Keith A Crandall, Dahiana Arcila, and Ricardo Betancur-R. Exon

probe sets and bioinformatics pipelines for all levels of fish phylogenomics. *Molecular Ecology Resources*, 21(3):816–833, 2021.

[27] Ron I Eytan, Benjamin R Evans, Alex Dornburg, Alan R Lemmon, Emily Moriarty Lemmon, Peter C Wainwright, and Thomas J Near. Are 100 enough? inferring acanthomorph teleost phylogeny using Anchored Hybrid Enrichment. *BMC evolutionary biology*, 15(1):1–20, 2015.

[28] Frank E Zachos. Mammalian phylogenetics: A short overview of recent advances. *Mammals of Europe-Past, Present, and Future*, pages 31–48, 2020.

[29] James E Tarver, Mario Dos Reis, Siavash Mirarab, Raymond J Moran, Sean Parker, Joseph E O'Reilly, Benjamin L King, Mary J O'Connell, Robert J Asher, Tandy Warnow, et al. The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome biology and evolution*, 8(2):330–344, 2016.

[30] Jacob A Esselstyn, Carl H Oliveros, Mark T Swanson, and Brant C Faircloth. Investigating difficult nodes in the placental mammal tree with expanded taxon sampling and thousands of ultraconserved elements. *Genome Biology and Evolution*, 9(9):2308–2321, 2017.

[31] Nathan S Upham, Jacob A Esselstyn, and Walter Jetz. Inferring the mammal tree: species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS biology*, 17(12):e3000494, 2019.

[32] Mark T Swanson, Carl H Oliveros, and Jacob A Esselstyn. A phylogenomic rodent tree reveals the repeated evolution of masseter architectures. *Proceedings of the Royal Society B*, 286(1902):20190672, 2019.

[33] Xiaoping Wang, Burton K Lim, Nelson Ting, Jingyang Hu, Yunpeng Liang, Christian Roos, and Li Yu. Reconstructing the phylogeny of new world monkeys (platyrrhini): evidence from multiple non-coding loci. *Current zoology*, 65(5):579–588, 2019.

[34] Carlos G Schrago and Hector N Seuánez. Large ancestral effective population size explains the difficult phylogenetic placement of owl monkeys. *American journal of primatology*, 81(3):e22955, 2019.

[35] Dan Vanderpool, Bui Quang Minh, Robert Lanfear, Daniel Hughes, Shwetha Murali, R Alan Harris, Muthuswamy Raveendran, Donna M Muzny, Mark S Hibbins, Robert J Williamson, et al. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLoS biology*, 18(12):e3000954, 2020.