NJMerge: A generic technique for scaling phylogeny estimation methods and its application to species trees

Supplementary Materials

Erin K. Molloy[1[0000−0001−5553−3312]] and Tandy Warnow[1[0000−0001−7717−3514]]

Department of Computer Science, University of Illinois at Urbana-Champaign, 201 N Goodwin Ave, Urbana, IL 61801
{emolloy2, warnow}@illinois.edu

Empirical Statistics on Simulated Datasets

Table S1: We report properties of simulated datasets. Average distance is the normalized Robinson-Foulds or RF [9] distance between the true species tree and the true gene tree, averaged across all 1000 genes in a replicate dataset. Gene tree estimation error is the normalized RF distance between the true and the estimated gene trees, averaged across all 1000 genes in a replicate dataset. Total gene tree discord is the normalized RF distance between the true species tree and the estimated gene tree, averaged across all 1000 genes in a replicate dataset. Values are the mean (± standard deviation) across 20 replicate datasets.

Number of Taxa	Sequence Type	Average Distance	Gene Tree Estimation Error	Total Gene Tree Discord
Moderate ILS (species tree height = 10M generations)				
100	exon	0.08 ± 0.02	0.38 ± 0.06	0.39 ± 0.06
100	intron	0.08 ± 0.02	0.26 ± 0.07	0.28 ± 0.06
1000	exon	0.10 ± 0.00	0.42 ± 0.04	0.43 ± 0.04
1000	intron	0.10 ± 0.00	0.30 ± 0.05	0.32 ± 0.05
Very High ILS (species tree height = 500K generations)				
100	exon	0.68 ± 0.02	0.57 ± 0.07	0.78 ± 0.03
100	intron	0.68 ± 0.02	0.43 ± 0.10	0.74 ± 0.03
1000	exon	0.69 ± 0.01	0.64 ± 0.05	0.81 ± 0.02
1000	intron	0.69 ± 0.01	0.51 ± 0.07	0.76 ± 0.03

Commands

SimPhy [6] Version 1.0.2 was run as

```
simphy -rs 3000 -rl F:3000 -rg 1 -st F:[species tree height] \
       -si F:1 -sl F:[number of taxa] -sb F:0.0000001 -sp F:200000 \
       -hs LN:1.5,1 -hl LN:1.2,1 -hg LN:1.4,1 -su E:10000000 \
       -so F:1 -od 1 -v 3  -cs 293745 -o [output directory name]
```

where the number of taxa was 100 or 1000, the species tree height was 10,000,000 or 500,000 generations, and the effective population size was constant at 200,000.

FastTree [8] Version 2.1.10 (SSE3) was run as

```
FastTree -nt -gtr -quiet -log fasttree-$gene.log \
     [input alignment fasta file] > [output FastTree-2 tree file]
```

ASTRID [13] Version 1.4 was run as

```
ASTRID -i [input gene tree list file] \
       -c [output distance matrix] -o [temporary file]
```

FastME [5] Version 2.1.5 was run as

```
fastme -m N -i [input distance matrix] -o [output tree file]
```

ASTRAL [14] Version 5.6.1 was run as

```
java java -Xms3200M -Xmx32000M ASTRAL/Astral/astral.5.6.1.jar \
     -i [input gene tree list file] -o [output ASTRAL-III tree file]
```

SVDquartets [1, 2] (PAUP* [12] Version 4a161 64-bit Centos) was run as

```
echo "exe [input alignment nexus file]; svd nthreads=16
    evalQuartets=all qfile=[output quartet file] qformat=qmc;
    savetrees file=[output SVDquartets tree file] format=newick;" |
    paup4a161_centos64 -n
```

RAxML [10] Version 8.2.12 (with pThreads SSE3) was run as

```
raxmlHPC-PTHREADS-SSE3 -m GTRGAMMA -F -p [seed] \
    -n [output name] -s [input alignment file] -T 16
```

Note that the option `-j` (to write checkpoints) was included for the 1000-taxon datasets only.

NJMerge was run as

```
python njmerge.py \
    -t [input constraint tree 1] ... [input constraint tree N] \
    -m [input internode distance matrix] \
    -x [input taxon name map for internode distance matrix]
    -o [output NJMerge tree]
```

Normalized RF distances were computed using Dendropy [11] Version 4.3.0 as

```
n1 = len(t1.internal_edges(exclude_seed_edge=True))
n2 = len(t2.internal_edges(exclude_seed_edge=True))
[fp, fn] = false_positives_and_negatives(t1, t2)
rf = float(fp + fn) / (n1 + n2)
```

where `t1` and `t2` are Dendropy tree objects.

INDELible Simulation

Like the simulation in [7], GTR+Γ model parameters (base frequencies, substitution rates, and alpha) were drawn from distributions. However, unlike the simulation in [7], we estimated separate distributions for exons, introns, and ultra-conserved elements (UCEs) from the Avian Phylogenomics Dataset [4]. We ran INDELible [3] Version 1.03 using custom Python scripts available on the Illinois Data Bank using the exon, intron, and UCE parameter distributions for genes 1-1000, 1001-2000, and 2001-3000, respectively. Only the exon-like and intron-like genes were used in this study due to limitations in computational resources. Data were simulated for this study using the protocol presented in [7].

Table S2: We report the distributions from which GTR+Γ model parameters were drawn to simulate sequences with INDELible.

Sequence Type	GTR Base Frequencies	GTR Substitution Rates	Gamma Parameter α
Exon	Dirchlet(79,57,60,53)	Dirchlet(3,9,2,4,11,7)	4.2
Intron	Dirchlet(55,38,43,63)	Dirchlet(37,133,20,48,120,29)	0.4
UCE	Dirchlet(68,45,45,68)	Dirchlet(19,66,10,27,67,19)	1.0

Results

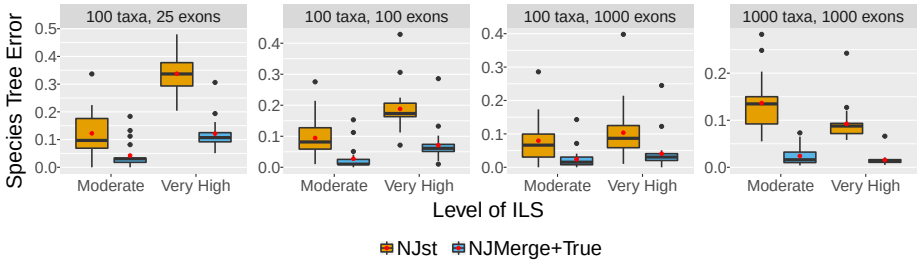
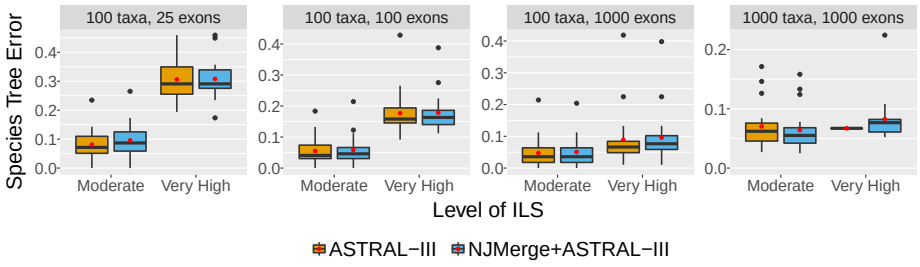
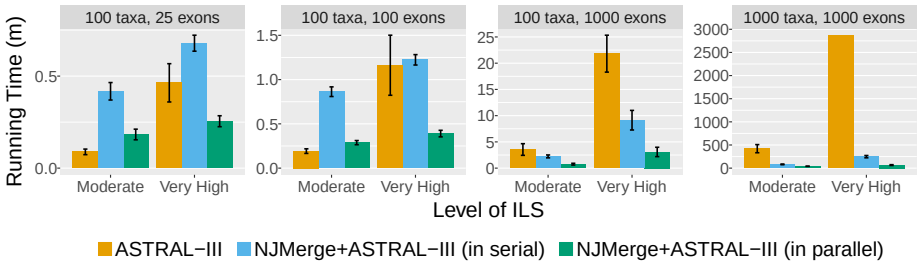


Fig. S1: Comparison of NJst (i.e., NJMerge without any subset trees) and NJMerge+True (i.e., NJMerge given subset trees defined by the true species tree as input) on exon-like datasets. Both NJst and NJMerge were run on the internode distance matrix computed using estimated gene trees. Species tree estimation error is defined as the normalized RF distance between true and estimated species trees. Bars show medians and red dots show means across replicate datasets. Box plots are defined by quartiles, extending from the first to the third quartiles. Whiskers extend to plus or minus 1.5 times the interquartile distance (unless greater/less than the maximum/minimum value).

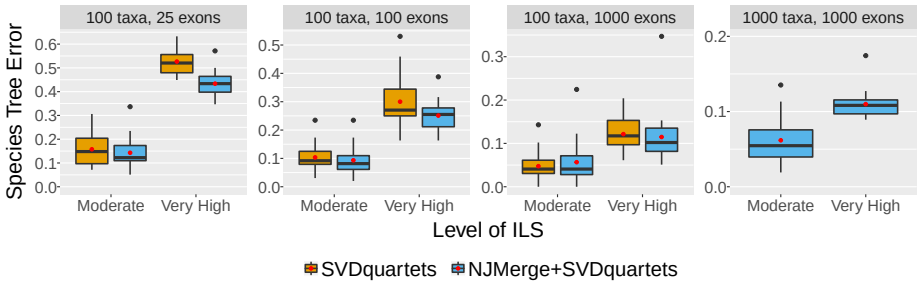


(a)

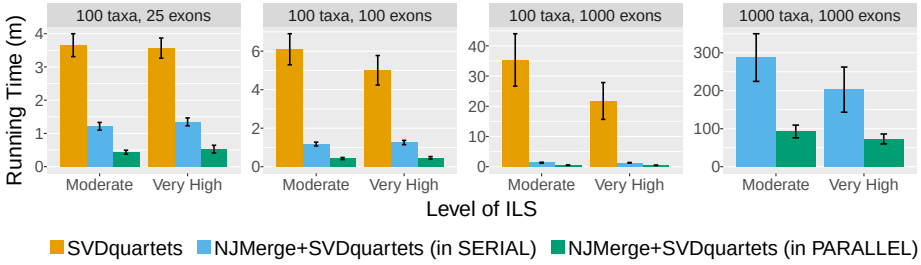


(b)

Fig. S2: Comparison of ASTRAL and NJMerge+ASTRAL (i.e., NJMerge given the ASTRAL subset trees as input) on exon-like datasets. Subplot (a) shows species tree estimation error (defined as the normalized RF distance between true and estimated species trees); bars represent medians and red dots represent means across replicate datasets. Box plots are defined by quartiles, extending from the first to the third quartiles. Whiskers extend to plus/minus 1.5 times the interquartile distance (unless greater/less than the maximum/minimum value). Subplot (b) shows running time (in minutes); bars represent means and error bars represent standard deviations, across replicate datasets. For NJMerge+ASTRAL, "in serial" or "in parallel" refers to whether subset trees could be estimated in serial or in parallel; see Equations (1) and (2) in the main text for more information. ASTRAL did not complete during the maximum wall-clock time of 48 hours on 19 out of the 20 replicate datasets with 1000 taxa and very high ILS.

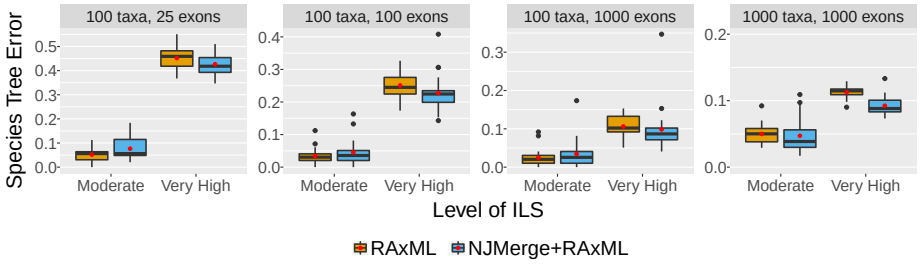


(a)

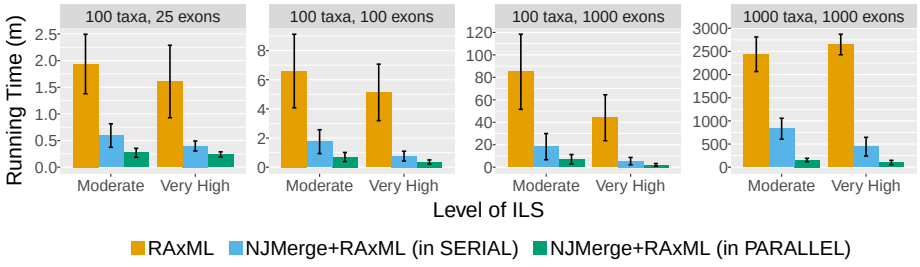


(b)

Fig. S3: Comparison of SVDquartets and NJMerge+SVDquartets (i.e., NJMerge given the SVDquartets subset trees as input) on exon-like datasets. Subplot (a) shows species tree estimation error (defined as the normalized RF distance between true and estimated species trees); bars represent medians and red dots represent means across replicate datasets. Box plots are defined by quartiles, extending from the first to the third quartiles. Whiskers extend to plus/minus 1.5 times the interquartile distance (unless greater/less than the maximum/minimum value). Subplot (b) shows running time (in minutes); bars represent means and error bars represent standard deviations, across replicate datasets. For NJMerge+SVDquartets, "in serial" or "in parallel" refers to whether subset trees could be estimated in serial or in parallel; see Equations (1) and (2) in the main text for more information. SVDquartets did not run on any datasets with 1000 taxa due to segmentation faults.



(a)



(b)

Fig. S4: Comparison of RAxML and NJMerge+RAxML (i.e., NJMerge given the RAxML subset trees as input) on exon-like datasets. Subplot (a) shows species tree estimation error (defined as the normalized RF distance between true and estimated species trees); bars represent medians and red dots represent means across replicate datasets. Box plots are defined by quartiles, extending from the first to the third quartiles. Whiskers extend to plus/minus 1.5 times the interquartile distance (unless greater/less than the maximum/minimum value). Subplot (b) shows running time (in minutes); bars represent means and error bars represent standard deviations, across replicate datasets. For NJMerge+RAxML, "in serial" or "in parallel" refers to whether subset trees could be estimated in serial or in parallel; see Equations (1) and (2) in the main text for more information. RAxML did not run on 3 of the 20 replicate datasets with 1000 taxa and moderate ILS due to "Out of Memory" errors. NJMerge+RAxML failed on one replicate datasets with 100 taxa, 25 genes, very high ILS, and exon-like sequences.

Table S3: Each species tree estimation method (ASTRAL-III, SVDquartets, or RAML) was run on the full dataset or on subsets in order to build constraint trees for NJMerge. We report the average (± standard deviation) species tree estimation error for (1) the tree produced by running a species tree method on the full set of species (all 100 or all 1000 taxa), (2) the tree produced by running a species tree method on subsets of species, (3) the tree produced by running NJst, and (4) running NJMerge using the same internode distance as NJst and the subsets from (2). Species tree estimation error (defined as normalized RF distance between the true and the estimated species tree) was averaged across 20 replicate datasets, unless the number of replicate datasets is otherwise noted in parentheses. When methods were run on subsets, species tree estimation error was averaged across all subsets and all replicate datasets. Note that the number of taxa in the subset trees was less than 30 for the 100-taxon datasets and less than 120 for the 1000-taxon datasets.

# Taxa	# Genes	Species Tree Height	Data Type	Method on Full	Method on Subset	NJst	NJMerge+Method
Method = ASTRAL-III							
100	25	10M	exon	0.08	0.07	0.12	0.10
100	25	10M	intron	0.06	0.05	0.07	0.06
100	25	500K	exon	0.31	0.25	0.34	0.31
100	25	500K	intron	0.26	0.20	0.30	0.26
100	100	10M	exon	0.06	0.04	0.09	0.06
100	100	10M	intron	0.04	0.03	0.05	0.04
100	100	500K	exon	0.18	0.12	0.19	0.18
100	100	500K	intron	0.12	0.09	0.14	0.13
100	1000	10M	exon	0.05	0.03	0.08	0.05
100	1000	10M	intron	0.02	0.02	0.04	0.03
100	1000	500K	exon	0.09	0.06	0.10	0.10
100	1000	500K	intron	0.05	0.04	0.06	0.06
1000	1000	10M	exon	0.07	0.04	0.14	0.06
1000	1000	10M	intron	0.05	0.03	0.11	0.05
1000	1000	500K	exon	0.07 (1)	0.07	0.09	0.08
1000	1000	500K	intron	0.06 (16)	0.05	0.07	0.06
Method = SVDquartets							
100	25	10M	exon	0.16	0.12	0.12	0.14
100	25	10M	intron	0.12	0.09	0.07	0.10
100	25	500K	exon	0.53	0.38	0.34	0.43
100	25	500K	intron	0.50	0.34	0.30	0.39
100	100	10M	exon	0.10	0.08	0.09	0.09
100	100	10M	intron	0.07	0.06	0.05	0.07
100	100	500K	exon	0.30	0.21	0.19	0.25
100	100	500K	intron	0.27	0.20	0.14	0.23
100	1000	10M	exon	0.05	0.04	0.08	0.06
100	1000	10M	intron	0.03	0.03	0.04	0.03
100	1000	500K	exon	0.12	0.08	0.10	0.11
100	1000	500K	intron	0.11	0.08	0.06	0.10
1000	1000	10M	exon	NA (0)	0.04	0.14	0.06
1000	1000	10M	intron	NA (0)	0.03	0.11	0.05
1000	1000	500K	exon	NA (0)	0.10	0.09	0.11
1000	1000	500K	intron	NA (0)	0.09	0.07	0.10
Method = RAxML							
100	25	10M	exon	0.05	0.04	0.12	0.08
100	25	10M	intron	0.04	0.03	0.07	0.05
100	25	500K	exon	0.45	0.38	0.34	0.43 (19)
100	25	500K	intron	0.42	0.34	0.30	0.39
100	100	10M	exon	0.03	0.02	0.09	0.05
100	100	10M	intron	0.02	0.02	0.05	0.03
100	100	500K	exon	0.25	0.18	0.19	0.23
100	100	500K	intron	0.24	0.18	0.14	0.21
100	1000	10M	exon	0.02	0.01	0.08	0.03
100	1000	10M	intron	0.02	0.01	0.04	0.02
100	1000	500K	exon	0.11	0.06	0.10	0.10
100	1000	500K	intron	0.09	0.06	0.06	0.08
1000	1000	10M	exon	0.05 (17)	0.02	0.14	0.05
1000	1000	10M	intron	NA (0)	0.01	0.11	0.03
1000	1000	500K	exon	0.11	0.08	0.09	0.09
1000	1000	500K	intron	0.14 (1)	0.08	0.07	0.08

Table S4: Each species tree estimation method (ASTRAL-III, SVDquartets, or RAML) was run on the full dataset (all 100 or all 1000 taxa) or on subsets in order to build constraint trees for NJMerge. We report the average running time ($\pm$ the standard deviation) in seconds across 20 replicate datasets, unless the number of replicate datasets is otherwise noted in parentheses. When methods were run on subsets, the time was measured per subset, and then average was taken across all subsets for all replicate datasets. Note that the 100-taxon datasets were decomposed into 4-6 subsets with a maximum subset size of 30 taxa, and the 1000-taxon datasets were decomposed into 10-15 subsets with a maximum subset size of 120 taxa.

# Taxa	# Genes	Species Tree Height	Data Type	Method on Full (s)	Method on Subset (s)	NJMerge (s)
Method = ASTRAL-III						
100	25	10M	exon	5 ± 1	4 ± 1	5 ± 2
100	25	10M	intron	5 ± 1	4 ± 1	5 ± 1
100	25	500K	exon	28 ± 6	8 ± 2	5 ± 1
100	25	500K	intron	22 ± 6	7 ± 2	6 ± 2
100	100	10M	exon	12 ± 2	10 ± 2	5 ± 1
100	100	10M	intron	9 ± 1	9 ± 2	5 ± 1
100	100	500K	exon	70 ± 20	15 ± 3	5 ± 1
100	100	500K	intron	50 ± 12	14 ± 3	5 ± 1
100	1000	10M	exon	213 ± 65	28 ± 10	5 ± 1
100	1000	10M	intron	95 ± 48	21 ± 5	5 ± 1
100	1000	500K	exon	1309 ± 206	121 ± 56	5 ± 1
100	1000	500K	intron	1096 ± 193	103 ± 57	5 ± 1
1000	1000	10M	exon	25231 ± 5154	239 ± 119	1939 ± 66
1000	1000	10M	intron	10545 ± 3823	126 ± 68	1939 ± 74
1000	1000	500K	exon	172346 (1)	1073 ± 529	1950 ± 283
1000	1000	500K	intron	149146 ± 14657 (16)	907 ± 394	1879 ± 24
Method = SVDquartets						
100	25	10M	exon	219 ± 20	15 ± 5	6 ± 2
100	25	10M	intron	257 ± 28	15 ± 5	5 ± 2
100	25	500K	exon	214 ± 18	17 ± 5	9 ± 6
100	25	500K	intron	238 ± 25	15 ± 5	8 ± 4
100	100	10M	exon	366 ± 47	14 ± 5	5 ± 1
100	100	10M	intron	528 ± 81	14 ± 5	6 ± 2
100	100	500K	exon	300 ± 45	15 ± 6	6 ± 2
100	100	500K	intron	403 ± 89	15 ± 6	6 ± 2
100	1000	10M	exon	2120 ± 507	16 ± 7	5 ± 2
100	1000	10M	intron	3817 ± 821	19 ± 9	5 ± 2
100	1000	500K	exon	1305 ± 356	16 ± 6	5 ± 2
100	1000	500K	intron	2240 ± 806	16 ± 6	5 ± 2
1000	1000	10M	exon	NA (0)	1238 ± 1142	2005 ± 124
1000	1000	10M	intron	NA (0)	2219 ± 2019	1999 ± 184
1000	1000	500K	exon	NA (0)	839 ± 803	2057 ± 178
1000	1000	500K	intron	NA (0)	1550 ± 1615	1975 ± 76
Method = RAxML						
100	25	10M	exon	116 ± 33	7 ± 5	5 ± 2
100	25	10M	intron	157 ± 47	10 ± 7	5 ± 2
100	25	500K	exon	96 ± 40	4 ± 3	8 ± 2 (19)
100	25	500K	intron	158 ± 69	4 ± 3	8 ± 3
100	100	10M	exon	396 ± 148	22 ± 16	5 ± 2
100	100	10M	intron	604 ± 177	32 ± 25	5 ± 1
100	100	500K	exon	308 ± 113	9 ± 7	6 ± 2
100	100	500K	intron	496 ± 237	10 ± 9	5 ± 2
100	1000	10M	exon	5097 ± 1955	238 ± 202	5 ± 1
100	1000	10M	intron	8343 ± 2611	426 ± 401	5 ± 1
100	1000	500K	exon	2641 ± 1196	71 ± 60	5 ± 1
100	1000	500K	intron	5106 ± 2414	83 ± 99	5 ± 1
1000	1000	10M	exon	146329 ± 21692 (17)	3887 ± 2023	2055 ± 165
1000	1000	10M	intron	NA (0)	6496 ± 3226	2010 ± 174
1000	1000	500K	exon	158973 ± 12955	2037 ± 1554	2006 ± 179
1000	1000	500K	intron	169440 (1)	3976 ± 3203	1933 ± 86

References

1. Chifman, J., Kubatko, L.: Quartet Inference from SNP Data Under the Coalescent Model. Bioinformatics **30**(23), 3317–3324 (2014). https://doi.org/10.1093/bioinformatics/btu530
2. Chifman, J., Kubatko, L.: Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. Journal of Theoretical Biology **374**, 35–47 (2015). https://doi.org/https://doi.org/10.1016/j.jtbi.2015.03.006
3. Fletcher, W., Yang, Z.: INDELible: A Flexible Simulator of Biological Sequence Evolution. Molecular Biology and Evolution **26**(8), 1879–1888 (2009). https://doi.org/10.1093/molbev/msp098
4. Jarvis, E.D., Mirarab, S., et al.: Whole-genome analyses resolve early branches in the tree of life of modern birds. Science **346**(6215), 1320–1331 (2014). https://doi.org/10.1126/science.1253451
5. Lefort, V., Desper, R., Gascuel, O.: FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Molecular Biology and Evolution **32**(10), 2798–2800 (2015). https://doi.org/10.1093/molbev/msv150
6. Mallo, D., De Oliveira Martins, L., Posada, D.: SimPhy : Phylogenomic Simulation of Gene, Locus, and Species Trees. Systematic Biology **65**(2), 334–344 (2016). https://doi.org/10.1093/sysbio/syv082
7. Mirarab, S., Warnow, T.: ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics **31**(12), i44–i52 (2015). https://doi.org/10.1093/bioinformatics/btv234
8. Price, M.N., Dehal, P.S., Arkin, A.P.: FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE **5**(3), 1–10 (2010). https://doi.org/10.1371/journal.pone.0009490
9. Robinson, D., Foulds, L.: Comparison of phylogenetic trees. Mathematical Biosciences **53**(1), 131–147 (1981). https://doi.org/10.1016/0025-5564(81)90043-2
10. Stamatakis, A.: RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**(9), 1312–1313 (2014). https://doi.org/10.1093/bioinformatics/btu033
11. Sukumaran, J., Holder, M.T.: DendroPy: a Python library for phylogenetic computing. Bioinformatics **26**(12), 1569–1571 (2010). https://doi.org/10.1093/bioinformatics/btq228
12. Swofford, D.L.: PAUP* (*Phylogenetic Analysis Using PAUP), Version 4a161 (2018), http://phylosolutions.com/paup-test/
13. Vachaspati, P., Warnow, T.: ASTRID: Accurate Species TRees from Internode Distances. BMC Genomics **16**(10), S3 (2015). https://doi.org/10.1186/1471-2164-16-S10-S3
14. Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S.: Astral-iii: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics **19**(6), 153 (2018). https://doi.org/10.1186/s12859-018-2129-y