ASTRAL-II: coalescent-based species tree estimation for hundreds of species and thousands of genes (Supplementary Material)

Siavash Mirarab and Tandy Warnow

Contents

List of Figures

List of Tables

1 Supplementary figures and tables

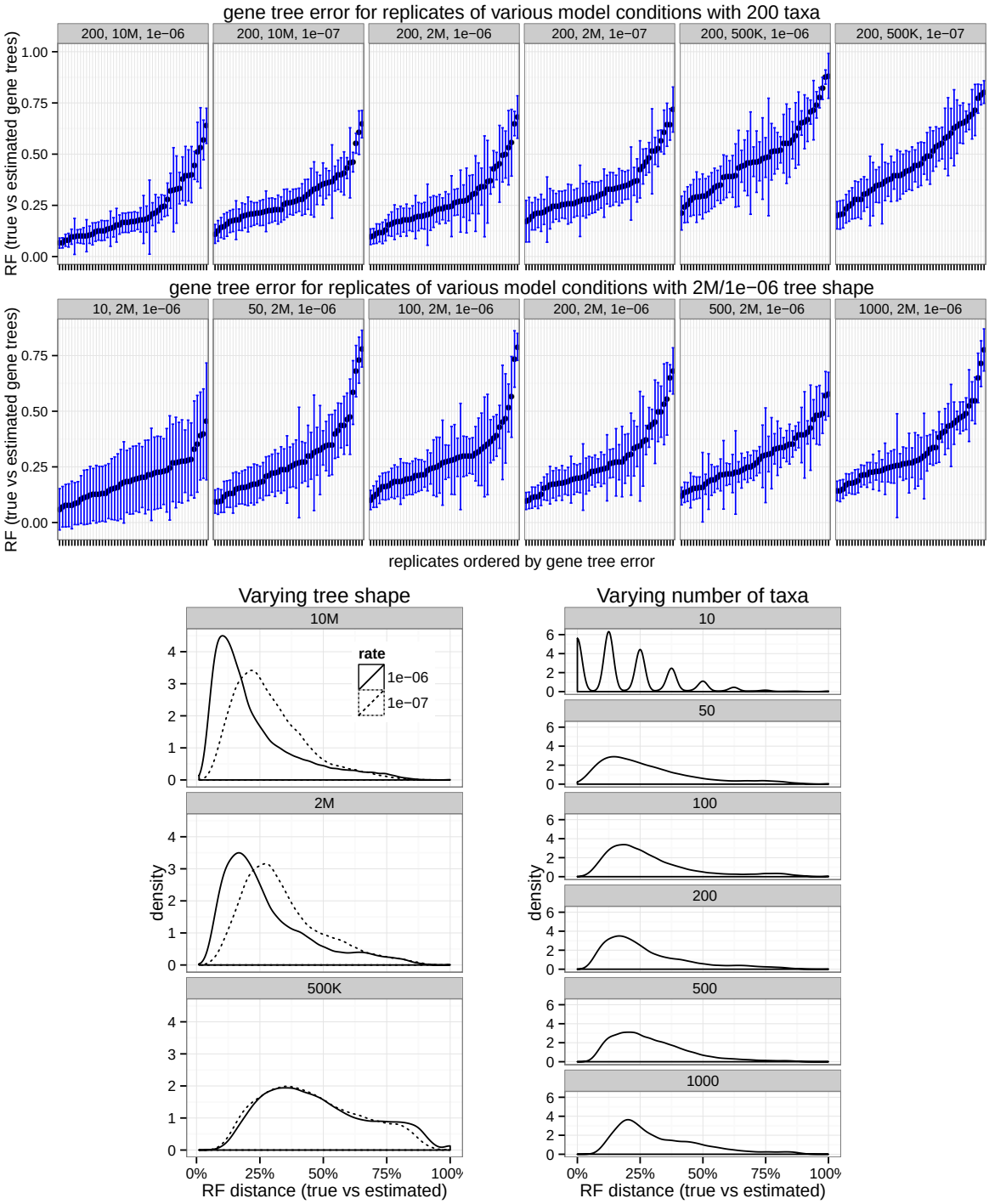


Figure S1: **Characteristics of the simulation - gene tree estimation error.** Many parameters (e.g. alignment length, gene tree length, and various substitution rates) were varied in a heterogenous way to simulate 50 replicates per model condition with varying gene tree estimation error. Top two panels: each box (box title: number of taxa, height, rate) shows averages and standard deviations of gene tree estimation error (across 1000 genes) for each replicate. Note wide variations in gene tree error across and within replicates. Bottom: both tree height and rate (left) affect the overall gene tree error, such that more ILS and deeper speciation both result in higher gene tree estimation error; when tree shape is fixed (2M, 1e-06), changing the number of taxa (right) has little impact on the gene tree estimation error.

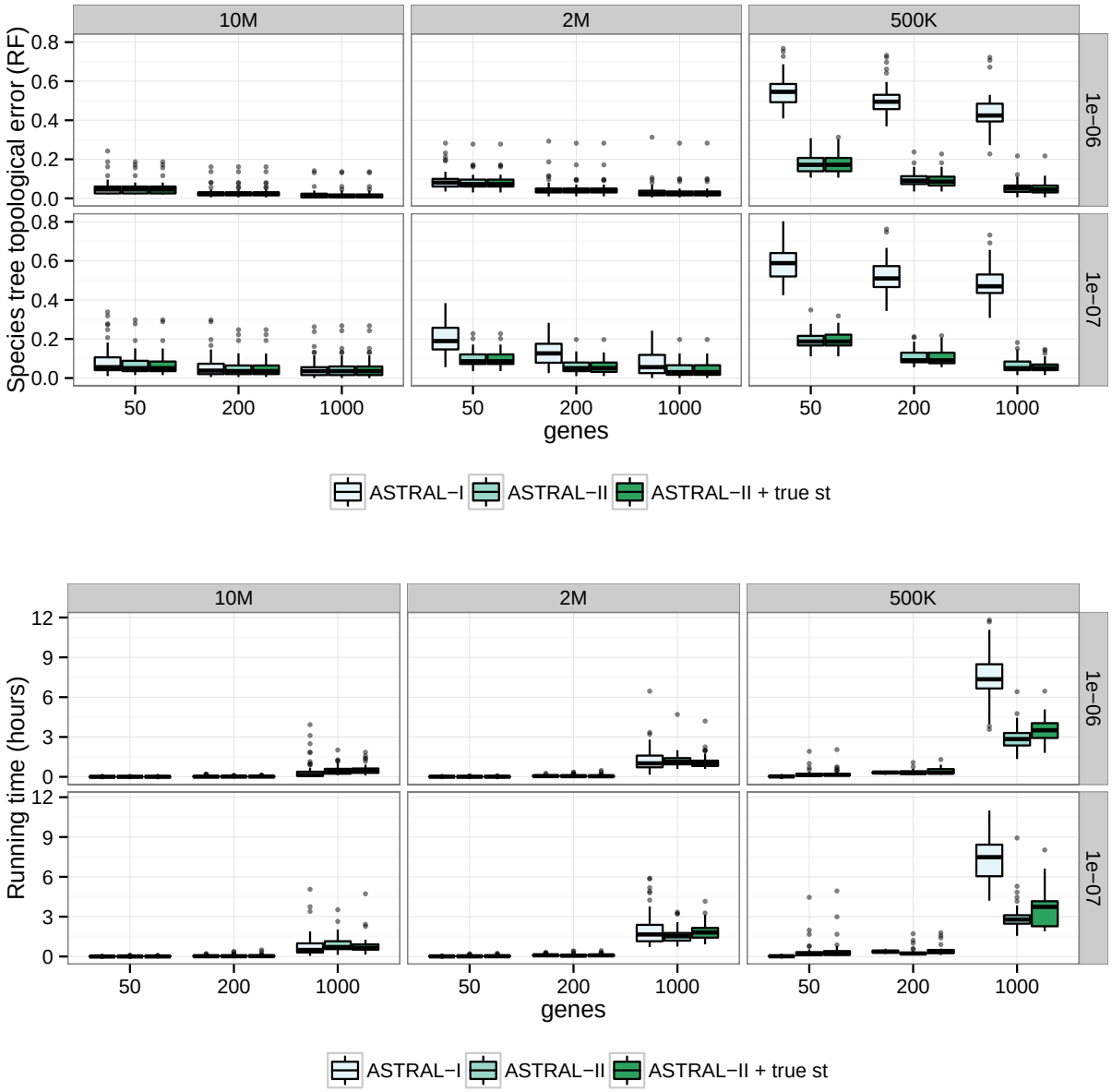


Figure S2: **Comparison of various variants of ASTRAL with 200 taxa and varying tree shapes and number of genes.** Species tree accuracy (top) and running times (bottom) are shown. ASTRAL-II + true st shows the case where the true species tree is added to the search space; this is included to approximate an ideal (e.g. exact) solution to the quartet problem.

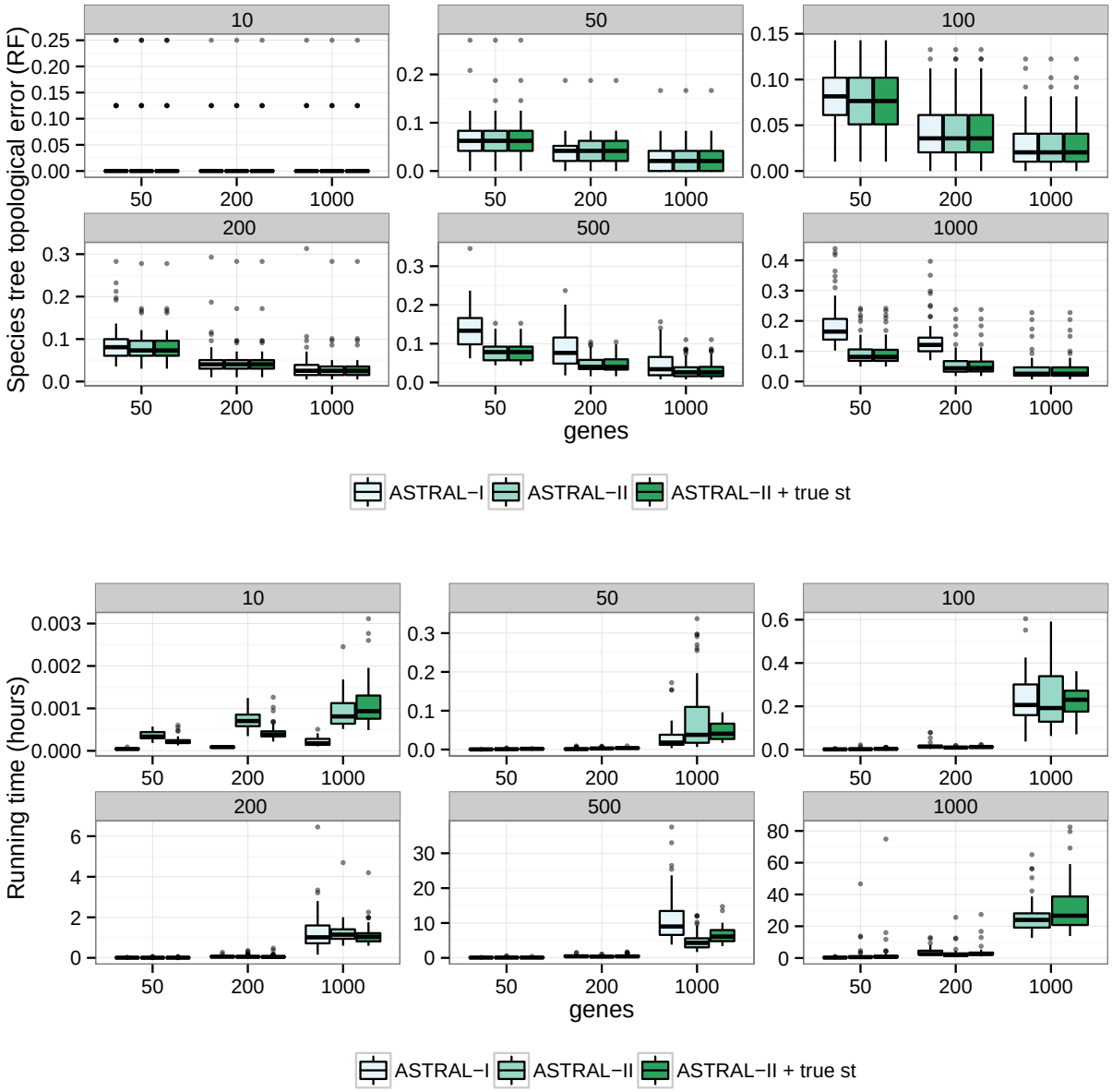


Figure S3: **Comparison of various variants of ASTRAL with varying number of taxa and genes (tree shaped fixed to 2M and 1e-06)**. Species tree accuracy (top) and running times (bottom) are shown. ASTRAL-II + true st shows the case where the true species tree is added to the search space; this is included to approximate an ideal (e.g. exact) solution to the quartet problem.

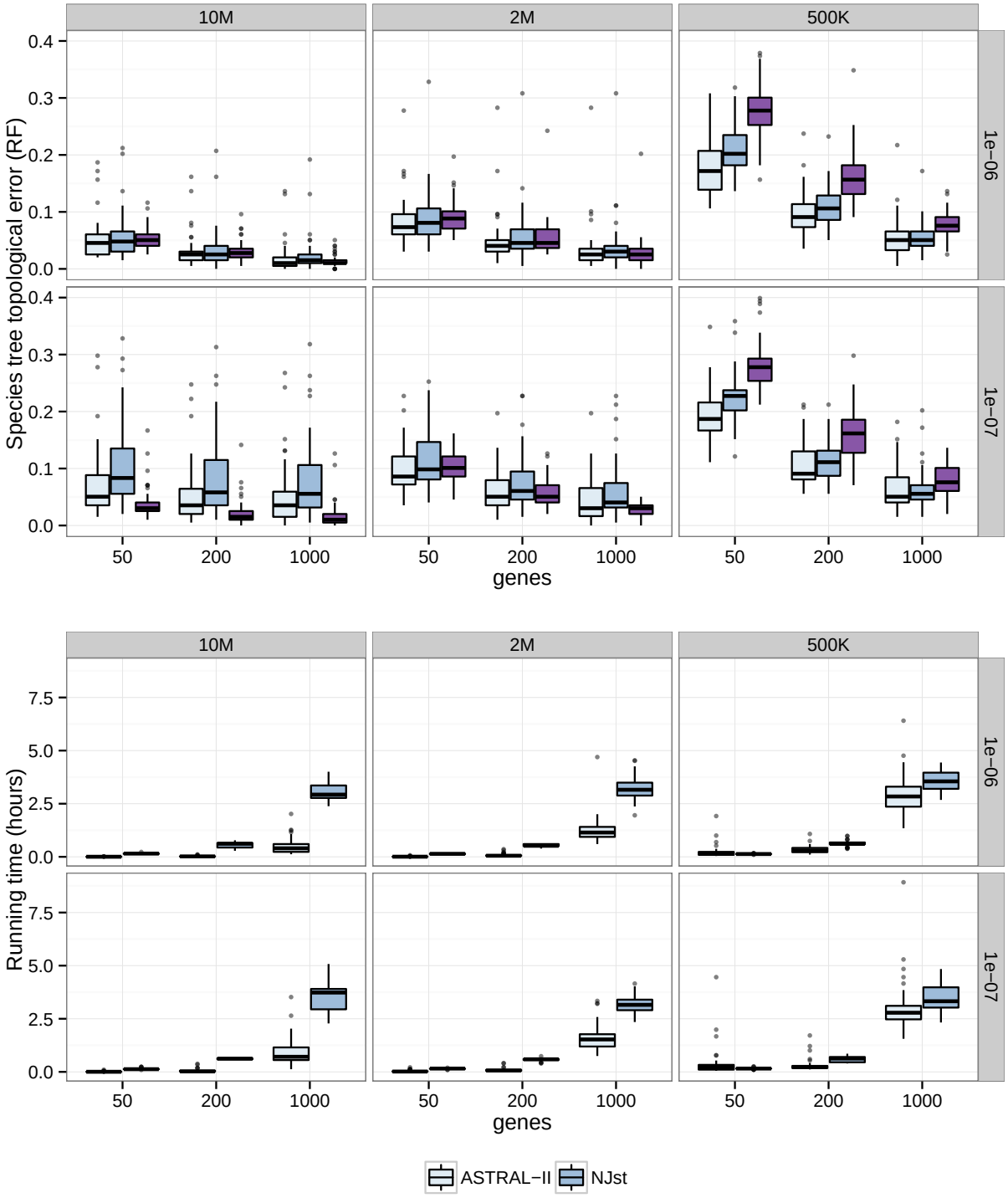


Figure S4: **Comparison of species tree accuracy with 200 taxa and varying tree shapes and number of genes**. Columns show varying tree lengths (with higher length corresponding to low ILS and lower length corresponding to higher ILS); rows show two different rates of speciation, which control whether speciation events tend to be close to the tips (1e-06) or close to the base (1e-07). Species tree accuracy (top) and running times (bottom) are shown.

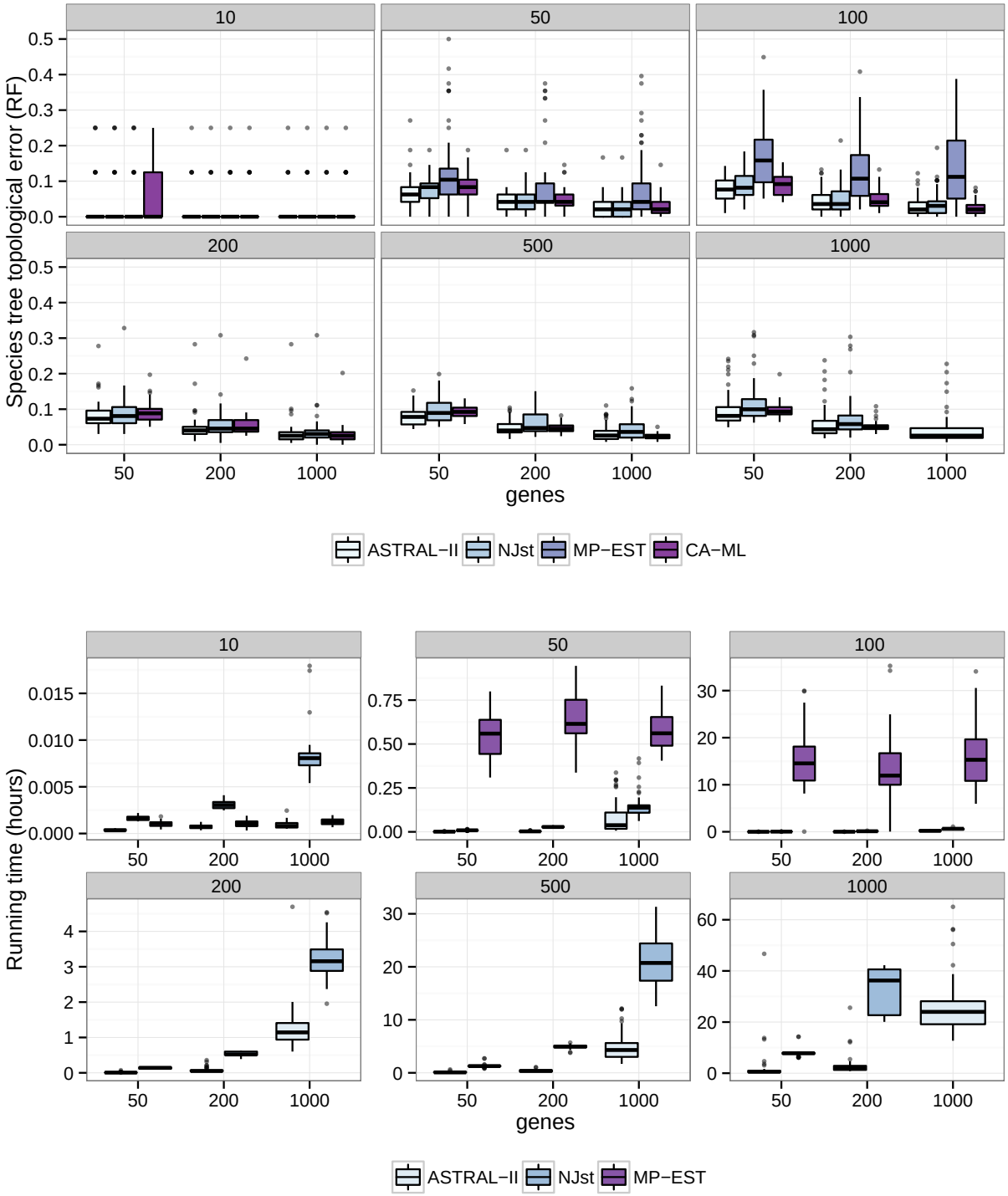


Figure S5: **Comparison of species tree accuracy with varying number of taxa and number of genes (tree shaped fixed to 2M and 1e-06)**. Boxes show varying number of taxa. Species tree accuracy (top) and running times (bottom) are shown. With 1000 genes of 1000 taxa, we were able to run only ASTRAL to completion.

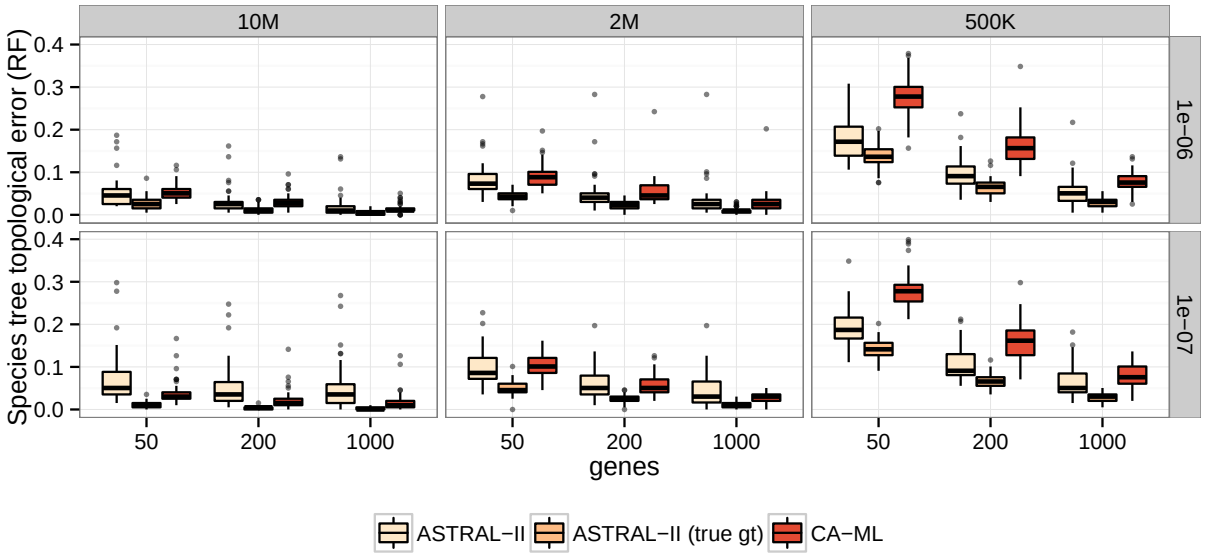


Figure S6: **Comparison of ASTRAL-II run on estimated and true gene trees with 200 taxa and varying tree shapes and number of genes.**

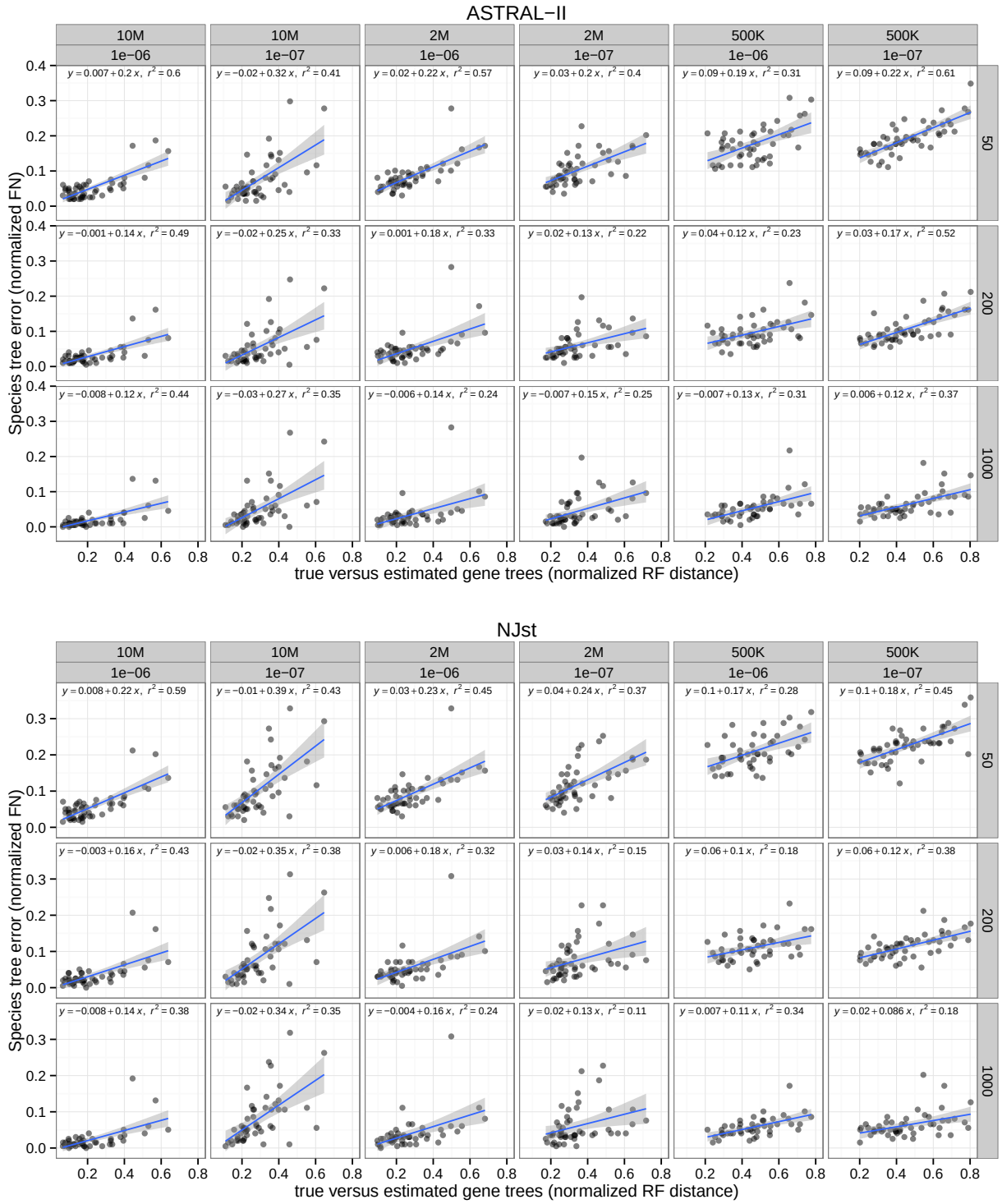


Figure S7: **Correlation between gene tree estimation error and species tree accuracy for ASTRAL and NJst with 200 taxa and varying tree shapes (columns) and number of genes (rows).** Gene tree and species tree error correlate well, and the correlation is stronger for fewer genes and *lower* levels of ILS.

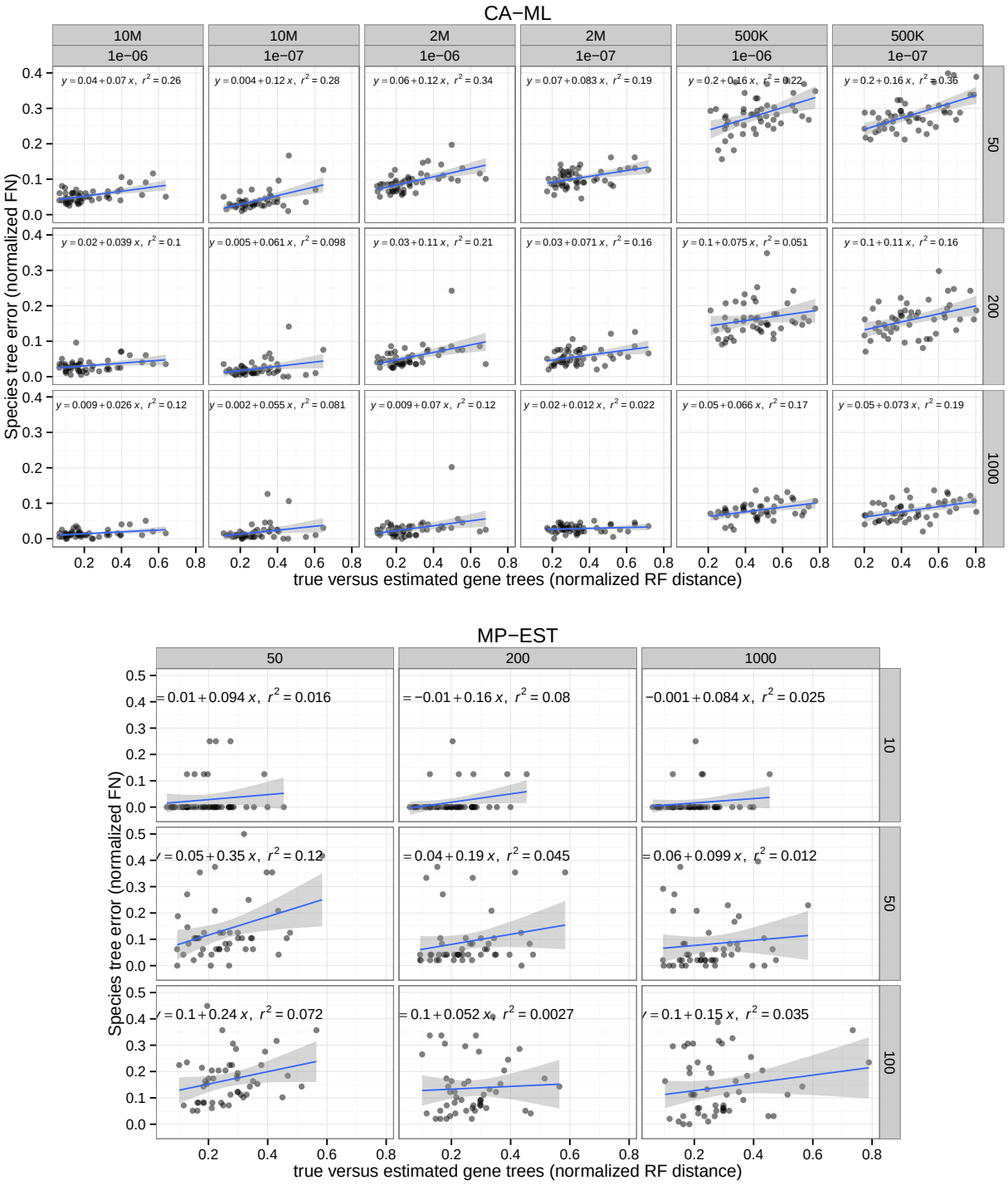


Figure S8: **Correlation between gene tree estimation error and species tree accuracy for CA-ML with 200 taxa and varying tree shapes (columns) and number of genes (rows) and MP-EST with varying number of genes and 50 taxa.** A correlation between gene tree error (controlled by parameters such as alignment length that also affect concatenation) and species tree error is detectable for concatenation also, but is smaller compared to NJst and ASTRAL. MP-EST also shows high levels of correlation between gene tree error and the species tree error.

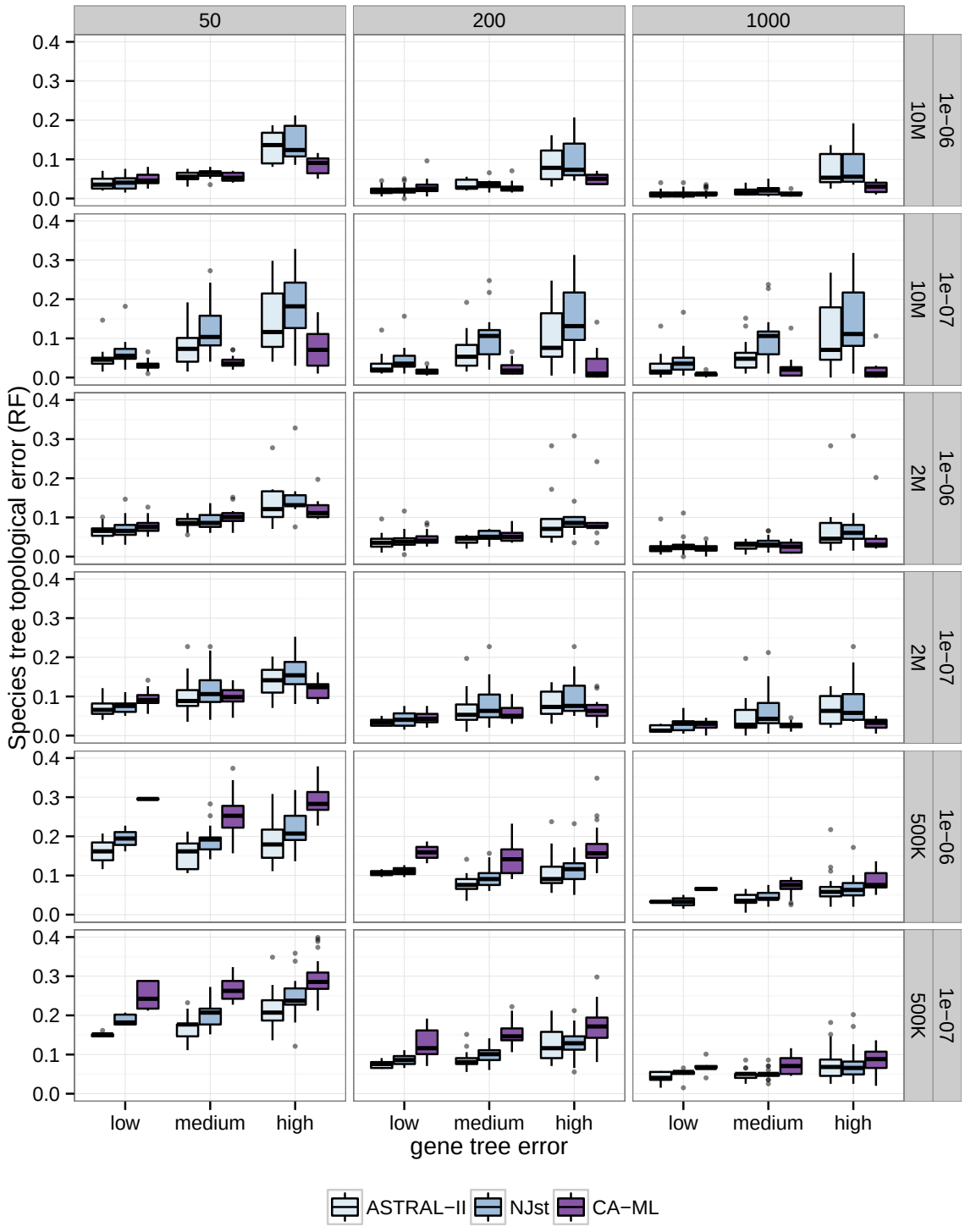


Figure S9: **Comparison of species tree accuracy with 200 taxa and varying tree shapes (rows), and varying number of genes (columns), divided into three categories of gene tree estimation error.**

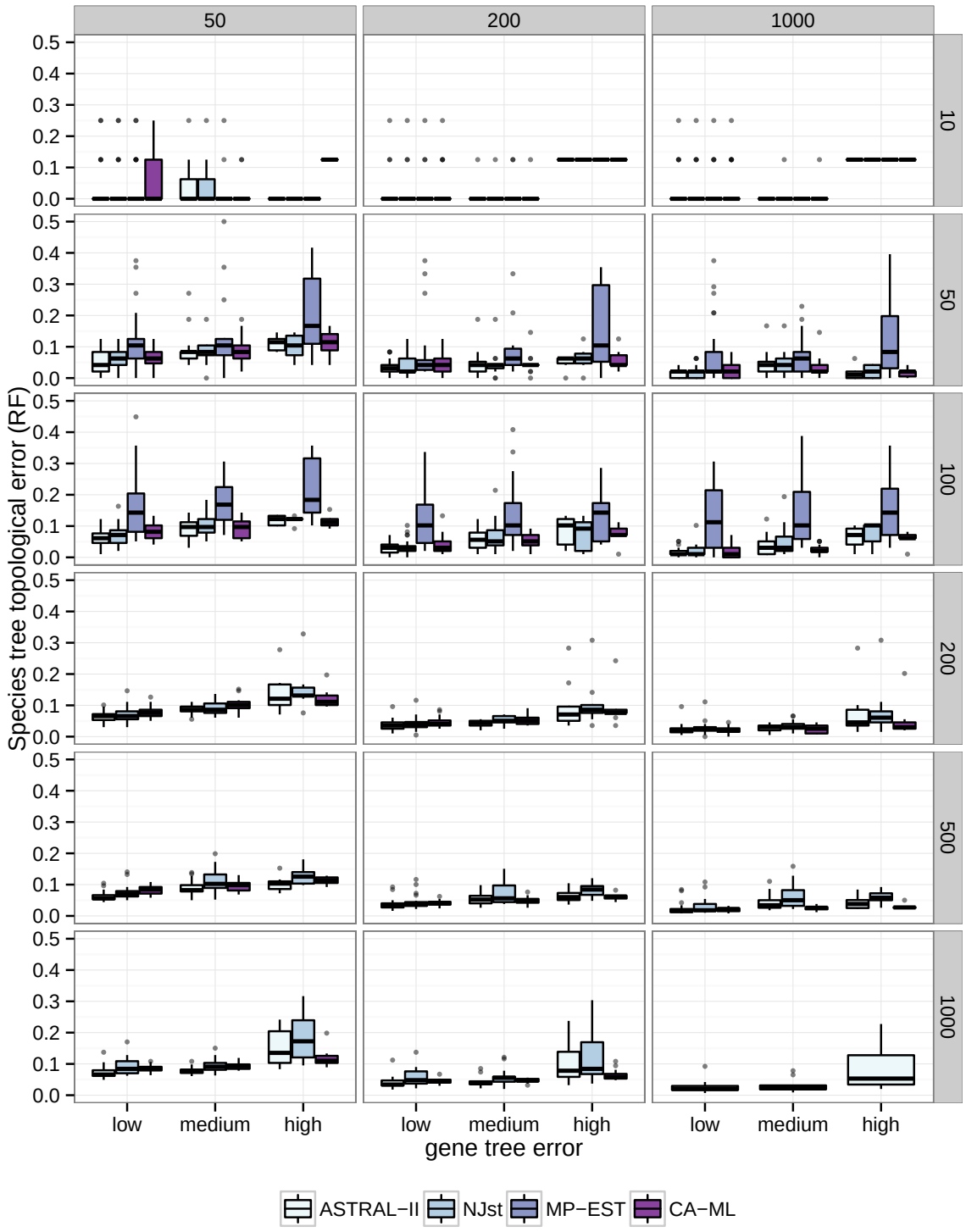


Figure S10: **Comparison of species tree accuracy with fixed tree shape (2M, 1e-06), varying number of taxa (rows), and varying number of genes (columns), divided into three categories of gene tree estimation error.**

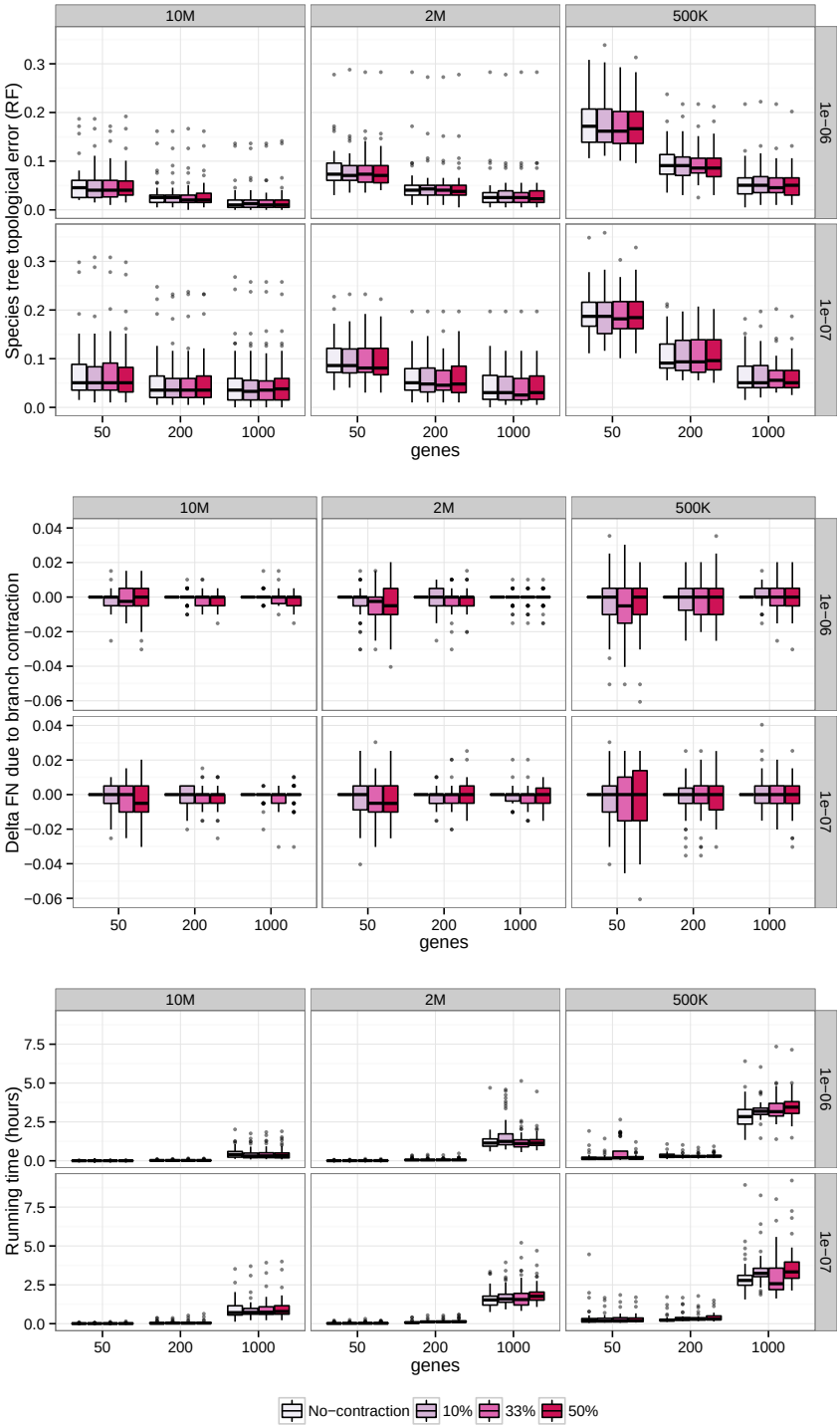


Figure S11: **Effect of contracting low support branches on ASTRAL with 200 taxa and varying tree shapes and number of genes.** Gene tree branches with FastTree SH-like local support below 10%, 33%, and 50% were contracted, and ASTRAL was run on these contracted gene trees. Species tree accuracy (top), change in species tree accuracy compared to the no-contraction ASTRAL tree (middle) and running times (bottom) are shown.

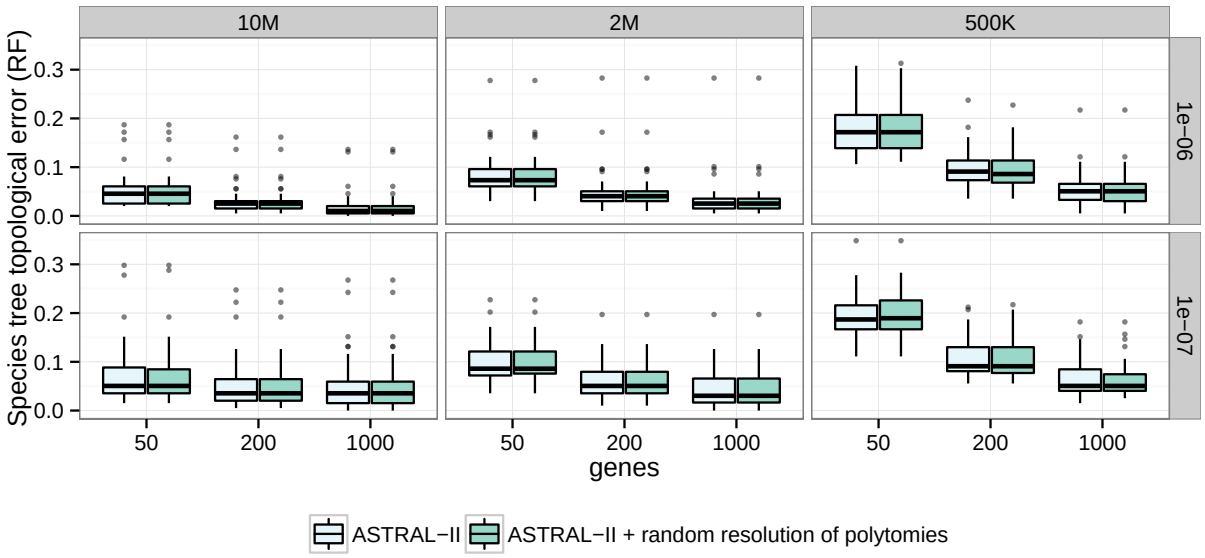


Figure S12: **Comparison of ASTRAL-II run on estimated gene trees with polytomies output by FastTree and with random resolutions of polytomies.** Results are with 200 taxa and varying tree shapes and number of genes.

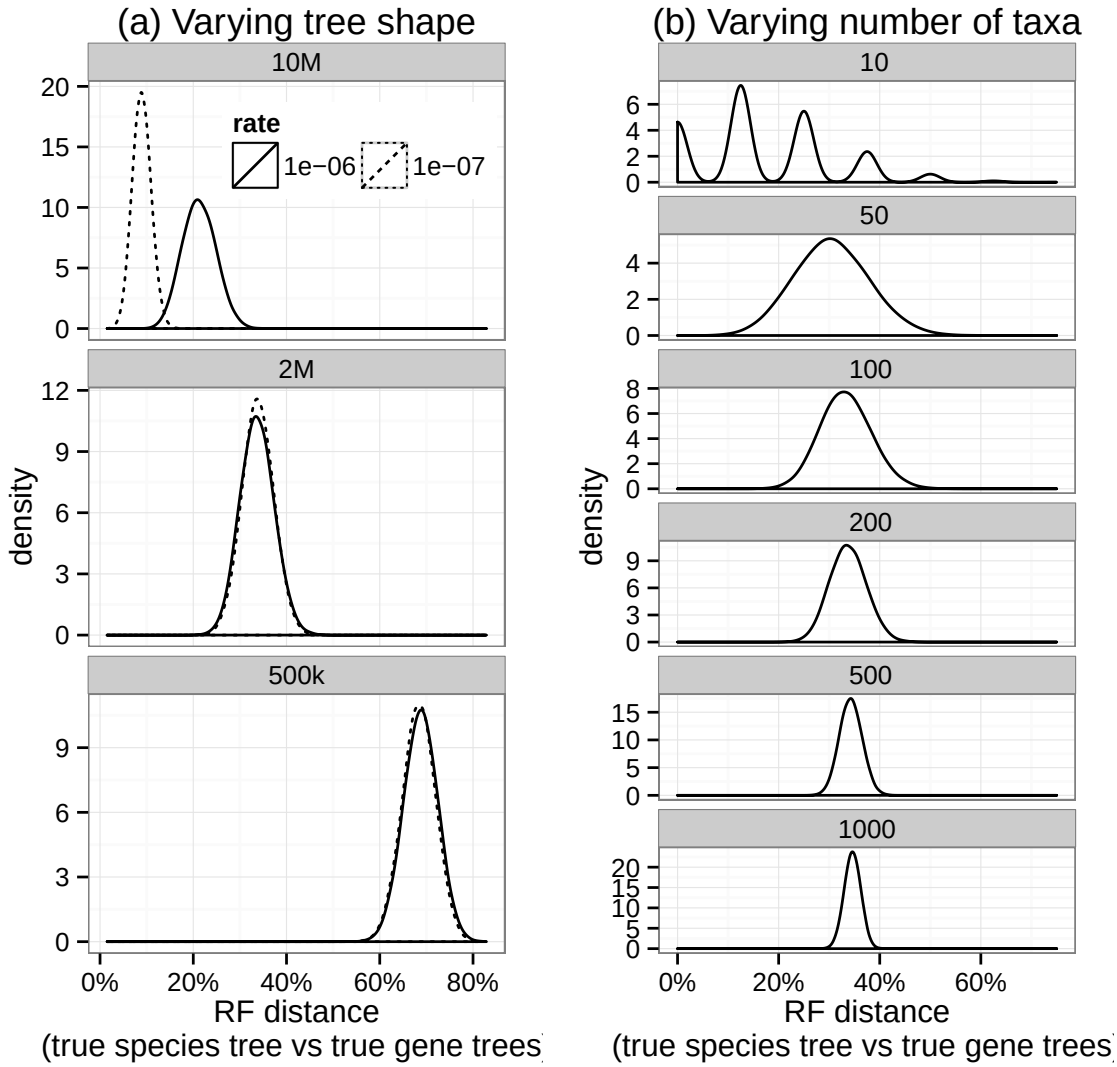
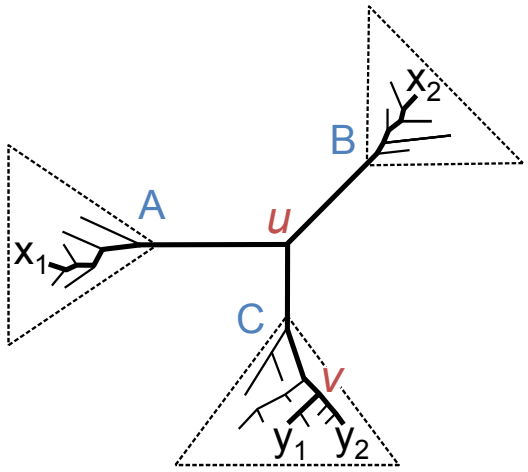
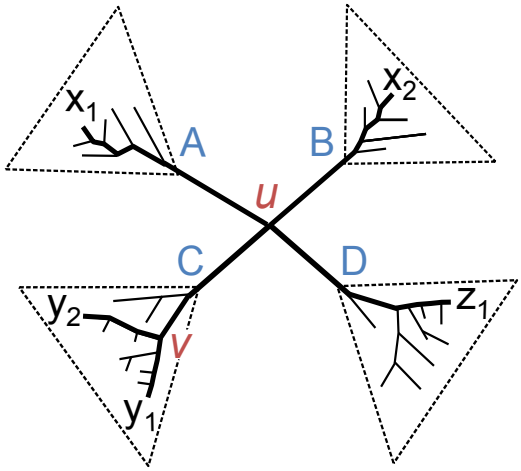


Figure S13: **Characteristics of the simulation - true gene tree discordance.** Density functions show RF distance between the true species tree and true gene trees (50 replicates of 1000 genes) for Dataset I (a) and Dataset II (b). Tree height directly affects the amount of true discordance; the speciation rate affects true gene tree discordance only with 10M tree length. The number of species has a relatively modest affect on the amount of true gene tree discordance, so that increasing the number of species with a fixed tree length results in somewhat shorter branches, and therefore more ILS. In particular datasets with 10 and 50 species have noticeably lower levels of discordance compared to other model conditions, but discordance does not seem to change substantially between 100 to 1000 species.



(a) tripartitions in fully resolved rooted trees.



(b) multi-partitions defined by a polytomy ($d = 4$).

Figure S14: **Tripartitions in unrooted gene trees.** (a) Each fully-resolved u in an unrooted tree defines a tripartition $(A|B|C)$ of the set of taxa, and conversely any given tripartition defines a node. Each induced quartet tree (e.g., $x_1x_2|y_1y_2$) maps to exactly two nodes in the tree. For example, the quartet tree on x_1, x_2, y_1, y_2 maps to u and v. Node u is where the paths from x_1 and x_2 to either y_1 or y_2 first join each other. Similarly, node v is where the paths from y_1 and y_2 to either x_1 or x_2 first join each other. Note that the number of quartets mapped to u is given by $\binom{|A|}{2}\binom{|B|}{1}\binom{|C|}{1} + \binom{|A|}{1}\binom{|B|}{2}\binom{|C|}{1} + \binom{|A|}{1}\binom{|B|}{1}\binom{|C|}{2} = \frac{|A||B||C|(|A|+|B|+|C|-3)}{2}$. Also note that any tree that includes the node u will induce all these quartet topologies that are mapped to node u. (b) A polytomy divides the set of taxa into more than three parts (here, we have $d = 4$ and therefore 4 parts). A quartet mapped to two nodes (e.g., $x_1x_2|y_1y_2$) is a resolved quartet topology and needs to be counted towards WQ scores. A quartet mapped to only one node (e.g., $x_1x_2|y_1z_1$) is an unresolved quartet, and does not contribute to the WQ score; these need to be ignored. By treating the polytomy as a collection of $\binom{d}{3}$ tripartitions (in this case, $A|B|C$, $A|B|D$, $A|C|D$, and $B|C|D$), we ensure that all resolved quartet trees are counted and all unresolved quartet trees are left out. For example, here, $x_1x_2|y_1z_1$ would be counted only if we choose taxa from four different partitions, and therefore will not be counted in our collection of $\binom{d}{3}$ tripartitions.

Table S1: **Species tree error on Dataset I.** We show average and standard error of RF percentage. Note that ASTRAL-II is always more accurate than NJst. For each row, the lowest average error and those error values that have an overlapping standard error with the lowest error value are in bold.

rate	height	genes	ASTRAL-II	NJst	CA-ML
1e-06	10M	50	**5.2±0.5**	**5.6±0.6**	**5.4±0.3**
1e-06	10M	200	**3.1±0.4**	**3.4±0.5**	**3.1±0.3**
1e-06	10M	1000	**2.0±0.4**	2.3±0.5	**1.4±0.2**
1e-06	2M	50	**8.4±0.6**	**9.1±0.7**	**9.2±0.4**
1e-06	2M	200	**5.0±0.6**	**5.6±0.6**	**5.5±0.5**
1e-06	2M	1000	**3.4±0.6**	3.9±0.6	**2.8±0.4**
1e-06	500K	50	**17.6±0.7**	20.9±0.7	27.9±0.7
1e-06	500K	200	**9.6±0.5**	11.0±0.5	16.2±0.7
1e-06	500K	1000	**5.3±0.5**	**5.7±0.4**	8.0±0.3
1e-07	10M	50	7.3±0.9	10.2±1.0	**4.0±0.4**
1e-07	10M	200	5.4±0.7	8.2±1.0	**2.2±0.3**
1e-07	10M	1000	5.0±0.8	8.0±1.0	**1.8±0.3**
1e-07	2M	50	**10.2±0.6**	11.7±0.7	**10.3±0.3**
1e-07	2M	200	**6.0±0.5**	7.5±0.7	**5.7±0.3**
1e-07	2M	1000	4.4±0.6	6.0±0.7	**2.8±0.2**
1e-07	500K	50	**19.3±0.7**	22.5±0.6	28.2±0.6
1e-07	500K	200	**10.7±0.6**	**11.4±0.5**	16.1±0.7
1e-07	500K	1000	**6.3±0.5**	**6.3±0.5**	8.0±0.4

Table S2: **Species tree error on Dataset II.** We show average and standard error of RF percentage. Note that ASTRAL-II is always more accurate than MP-EST, and more accurate than NJst under all conditions except one (50 taxa and 50 genes), where NJst is slightly more accurate (7.2% vs. 7.3%). For each row, the lowest average error and those error values that have an overlapping standard error with the lowest error value are in bold.

taxa	genes	ASTRAL-II	NJst	CA-ML	MP-EST
10	50	**2.8±1.0**	**2.8±1.0**	3.8±0.9	**2.8±1.0**
10	200	**1.5±0.7**	**1.5±0.7**	**1.8±0.7**	**1.8±0.7**
10	1000	**1.5±0.7**	**1.8±0.7**	**2.1±0.8**	**1.5±0.7**
50	50	**7.3±0.7**	**7.2±0.6**	**7.8±0.6**	13.5±1.7
50	200	**4.2±0.5**	**4.4±0.5**	**4.5±0.4**	9.1±1.5
50	1000	**2.6±0.4**	**2.7±0.5**	**2.7±0.4**	8.2±1.5
100	50	**7.9±0.5**	**8.7±0.5**	9.1±0.4	16.9±1.3
100	200	**4.8±0.5**	**5.1±0.6**	**4.7±0.4**	13.7±1.5
100	1000	**3.0±0.4**	3.9±0.6	**2.5±0.3**	14.1±1.55
200	50	**8.4±0.6**	**9.1±0.7**	**9.2±0.4**	
200	200	**5.0±0.6**	**5.6±0.6**	**5.5±0.5**	
200	1000	**3.4±0.6**	3.9±0.6	**2.8±0.4**	
500	50	**8.0±0.4**	9.7±0.5	9.2±0.3	
500	200	**4.9±0.3**	6.1±0.5	**4.7±0.2**	
500	1000	3.3±0.4	4.7±0.5	**2.3±0.1**	
1000	50	**9.9±0.7**	12.1±0.9	**9.8±0.3**	
1000	200	**6.0±0.7**	7.9±0.9	**5.1±0.2**	
1000	1000	**4.5±0.7**			

Table S3: **- Functions used in additions to X using greedy consensus (Algorithm 3).** A detailed description of various functions used in Algorithm 3 is given here.

Function	Description
$polytomies(gc)$	For a given unrooted tree gc, all nodes with degree $d > 3$ are returned.
$greedy(\mathcal{G}, t, b)$	Finds bipartitions in all input trees in $\mathcal{G}$ and for each bipartitions notes its frequency. Sorts bipartitions by the descending order of frequency (with arbitrary tiebreakers) and discards those with frequency below t. Starts with a fully unresolved tree (i.e., the star tree), and adds bipartitions one at a time according to the order; if a bipartition conflicts with the tree, ignores it. At the end, if b is true, any remaining polytomies in the tree are randomly resolved. The branches (i.e., bipartitions) in the resulting tree are labelled by their bipartition frequency (i.e., their frequency in trees in $\mathcal{G}$).
$updateX(t)$	Lists all bipartitions from tree t and adds them to the set X; notes which bipartitions are new and which are not. When edges in t have a frequency label (e.g., the labels generated by the *greedy* function), this function returns the maximum label of any *new* bipartition added to X.
$clusters(p)$	An unrooted node p with degree d divides taxa into d subsets (see Fig. S14b). This function returns the partitions defined by p.
$upgma(S, C)$	Runs the UPGMA algorithm using similarity matrix S on n taxa. By default, starts from n singleton clusters, one per taxa, and in each step, combines the two clusters with highest similarity. The similarity of two clusters is the average similarity between all pairs of leaves chosen each from one of the two clusters. When C is given, instead of starting with n singleton clusters, UPGMA starts by groups defined in C.
$randSample(p)$	Selects a random taxon from each partition defined by the node p.
$resolve(p, r)$	The input p is a node in an unrooted tree with leaf set L, and r is an unrooted tree on $L' \subset L$ such that L' contains exactly one leaf from each partition defined by p. Note that the tree r will be compatible with the tree that includes p. Every bipartition in r defines a further resolution of p. This function resolves p according to r and returns the results.
$pectinate(O)$	O is an ordered list of taxa. This function returns a pectinate unrooted tree based on O. For example, for $O = (a, d, e, c, b)$, the results is $(a, (d, (e, (c, b))))$.
$sortBy(S, s, sample)$	Sorts a list of taxa s based on their decreasing similarity to *sample* and according to the similarity matrix S.
Constants	$THS = \{0, \frac{1}{100}, \frac{1}{50}, \frac{1}{20}, \frac{1}{10}, \frac{1}{4}, \frac{1}{3}\}$; $MIT = 10$; $RWD = 2$; $FRQ = \frac{1}{100}$; $LTH = \frac{1}{100}$

2 Simulation

2.1 SimPhy Parameters

We used the following parameters in our simulation using SimPhy. The scripts for the simulation are given at `http://www.cs.utexas.edu/users/phylo/software/astral/`.

Table S4: Parameters used in SimPhy simulations.

Arg.	Description	Value	Notes
RS	number of replicates	50	
RL	number of loci	1000	
RG	number of genes	1	no duplications
ST	maximum tree length	500K, 2M, or 10M	
SI	number of individuals per species	1	
SL	number of leaves	10,50,100,200,500, or 1000	
SB	birth rates	0.000001, 0.0000001	
P	global population sizes	200000	
HS	Species-specific branch rate heterogeneity modifiers	Log normal (1.5,1)	
HL	Locus-specific rate heterogeneity modifiers	Log normal (1.2,1)	
HG	Gene-tree-branch-specific rate heterogeneity modifiers	Log normal (1.4,1)	
U	Global substitution rate	Exponential (10000000)	
SO	Outgroup branch length relative to half the tree length	1	
CS	Random number generator seed	293745	

2.2 Indelible Parameters

We used a perl script available also at `http://www.cs.utexas.edu/users/phylo/software/astral/` to draw parameters for the Indelible simulations. For each replicate, some hyperparameters are first drawn and these hyperparameters affect how the actual parameters are drawn for each gene in that replicate.

Gene Length: The alignments lengths are drawn from log normal distributions for genes of each replicate. For each replicate, a hyperparameter controls the two model parameters of the log normal distribution. The log mean is drawn uniformly between 5.7 and 7.3, which correspond to 300 sites to 1500 sites. Thus, the average alignment length for each replicate is a random value between 300 and 1500. The log standard deviation for the log normal distribution is also drawn uniformly between 0.0 and 0.3.

Base frequencies: We used a Dirichlet(36,26,28,32) to draw the base frequencies for A, C, G, and T. These values were calculated using maximum likelihood estimation form a collection of three large scale multi-locus datasets: 1KP dataset, Song et al Mammalian dataset, and Avian phylogenomics dataset. The base values used for this maximum likelihood estimation and the corresponding scripts are available at `http://www.cs.utexas.edu/~phylo/software/astral/`.

Substitution matrices: As with base frequencies, GTR matrices were drawn from a Dirichlet(16,3,5,5,6,15) and these parameters were also estimated using maximum likelihood from our empirical data.

Rates-across-sites shape parameter: α was drawn from an exponential distribution with rate 1.2, with values below 0.1 discarded. Like rates and base frequencies, these values were estimated from real data.