Originally published 12 December 2014; revised 2 January 2015

www.sciencemag.org/content/346/6215/1250463/suppl/DC1

Supplementary Materials for

Statistical binning enables an accurate coalescent-based estimation of the avian tree

Siavash Mirarab, Md. Shamsuzzoha Bayzid, Bastien Boussau, Tandy Warnow*

*Corresponding author. E-mail: warnow@illinois.edu

Published 12 December 2014, *Science* **346**, 1337 (2014)
DOI: 10.1126/science.1250463

This PDF file includes

Correction: The link http://www.cs.utexas.edu/users/phylo/datasets/binning/ was not working properly and has been replaced.

Supplementary Materials

Table of Contents

List of Figures

List of Tables

1 Materials and Methods

1.1 Binning technique

Our proposed statistical binning technique is applicable whenever estimated gene trees are used and the underlying sequence alignments are available. (Therefore, the binning technique does not make sense to use with true gene trees, since there are no alignments to combine.)

The statistical binning technique we present is parameterized by a minimum support threshold value, which we call t, set by default to 50% for any dataset with 1000 genes or more, or to 75% for datasets with less than 1000 genes. We first estimate individual gene trees using RAxML (20 independent runs) with branch support using 200 replicates of bootstrapping. We then contract branches in these gene trees that have support below threshold t. We compare all pairs of contracted gene trees in terms of tree compatibility, and note whether each pair of gene trees have any incompatible branches (two branches are incompatible if they cannot be both present in the same tree, i.e. if they are not combinable).

The test for pairwise compatibility is performed as follows. Let T_1 and T_2 be two gene trees in which all branches with support less than $t\%$ have already been contracted; these are treated as unrooted trees. Each branch in these trees defines a bipartition on the taxon set S, obtained by deleting the branch. Let e and e' be a pair of branches (one in each tree), with e defining bipartition (A_e, B_e) and e' defining bipartition $(A_{e'}, B_{e'})$. Then there is a tree T_3 containing both bipartitions if and only if one of the four pairwise intersections $A_e \cap A_{e'}$, $A_e \cap B_{e'}$, $B_e \cap A_{e'}$, and $B_e \cap B_{e'}$ is empty (*55*). The two trees T_1 and T_2 are compatible if all pairs of branches are compatible; otherwise they are incompatible. Testing two unrooted trees for compatibility can be computed in linear time (*55*).

The pairwise compatibility between gene trees is represented in an incompatibility graph, where each gene tree is represented as a vertex, and two vertices are connected by an edge if and only if their gene trees are incompatible. We formulate the binning problem as coloring the vertices of the incompatibility graph such that no adjacent nodes have the same color. Because there are no edges between any two vertices with the same color, a set of vertices with the same color defines a set of genes that have no incompatibility at or above the support threshold t.

A natural optimization problem is to find the minimum number of colors required for vertex coloring of a graph so that no two adjacent vertices have the same color. This is a well-studied problem in computer science that is known to be NP-hard, and so cannot currently be solved exactly for large graphs. However, there are heuristics for finding good vertex colorings (*56*). For our binning purpose, finding the absolute minimum number of bins is not crucial (or even necessarily desirable). Instead, it is important to avoid unbalanced bins (some very small, and some very large bins), for two reasons. The first, and most important reason, is statistical: methods like MP-EST that have statistical guarantees use the estimated distributions on gene trees to estimate the species tree. Thus, maintaining (or improving) the accuracy of the estimated distribution on gene trees is essential to statistical performance guarantees. The second reason is practical: unbalanced binning can create very small bins, and hence fail to benefit from binning, so that supergene trees will have poor accuracy. Thus, for both statistical and practical reasons, we wish to keep bin sizes approximately the same, while having as few bins as possible (maximizing the resolution and accuracy of supergene trees), and we do this taking combinability into account. Thus we seek to partition the genes into bins so that they all have approximately the same size, while not diverging too far from the minimum number of bins achievable through available heuristics. (In other words, we are willing to increase the number of bins, if this results in more balanced bins.) This is the *balanced vertex coloring* problem.

Our approach to achieve a balanced vertex coloring is based on a modification to the Brélaz heuristic algorithm (*56*), one of the most effective techniques for minimum vertex coloring. The Brélaz algorithm first finds a large clique in the graph, and creates a color for each vertex in the clique. It then orders the remaining vertices by a "saturation" measure (equal to the number of distinct colors at the neighbors of the vertex), and processes the vertices one by one in the descending order of saturation. For each vertex, the existing color classes (sets of vertices defined by a given color) are tested, and the vertex is added to the first set that has no conflict with it (note that adding a vertex to a color class means assigning it the color associated with the class). If no existing set can accommodate the vertex, a new color is created and assigned to that vertex. Each time a vertex is processed, the saturation scores of its neighbors are updated. Our modification to this heuristic is that for each vertex being processed, we first order the existing color classes based on their current size and examine them in ascending order. This ordering ensures that the new vertex is added to the smallest color class that it can be added to.

To implement the balanced vertex coloring heuristic, we modified a publicly available implementation of a modification to the Brélaz heuristic (we used an implementation by Shalin Shah available at http://shah.freeshell.org/graphcoloring/). Shah's algorithm provides extra features (an iterated Greedy and local search heuristic) to reduce the number of colors, which we removed so that only the Brélaz heuristic remained. We then modified the heuristic to use the balancing strategy (adding the new vertex according to the color class size). Note that the balanced vertex coloring heuristic produces very balanced bins over the course of the algorithm while also using a very small number of colors, very close to the number of colors achieved by the unmodified Brélaz heuristic, as shown in Figure S10.

Once the bins are formed, the alignments of the genes in each bin are concatenated to form supergene alignments, and supergene trees are estimated for each supergene using RAxML, again with 20 independent runs and 200 bootstrap replicates. The resulting bootstrap supergene trees are then used as input to a summary method (MP-EST, Greedy, MRP, etc.).

1.2 Simulation study protocol

We simulated two sets of datasets, one based on the avian dataset with 48 species and 14,446 loci studied in (*31*), and one based on the mammals dataset with 37 species and 447 loci studied in (*6*).

The avian dataset had exons from 8251 genes, introns from 2516 of these, and 3679 ultraconserved elements (UCEs) with their flanking sequences, totalling 14,446 different genomic regions. RAxML under GTRGAMMA model was used to compute bootstrapped gene trees for each of these loci. This analysis showed that the average bootstrap support (BS) within the different partitions varied substantially, with exons having 24% BS, UCEs having 39%, and introns having 48%. The longest introns (with at least 10,000 nucleotides) had the best average bootstrap support of 59%. The average bootstrap support among the gene trees was therefore quite low (only 32%), since it was largely influenced by the exons, which contributed the largest number of gene trees.

The mammalian dataset had 37 species and 447 loci, with average bootstrap support of 71%. These loci were selected based on stringent criteria of orthology identification, and were a restricted version of a larger collection of markers. Song *et al.* first identified more than 700 potential orthologs and then filtered the set to these 447 genes based on various criteria, such as requiring that all species be present in all the loci (*6*).

We created model conditions that resembled the again and mammalian biological datasets in terms of average bootstrap support per gene and total number of genes. We also explored analyses with different numbers of gene trees of varying bootstrap support.

For the mammalian simulation, we used the unbinned MP-EST tree (see Fig. S19) as the model tree. Note that this tree had branch lengths in coalescent units, and thus all the necessary information for simulation under the multi-species coalescent process. For simulating the avian dataset, we used the binned MP-EST tree on the TENT matrix from (*31*), but re-estimated the branch lengths on that model tree using only the longest genes with at least 10,000 sites. This was necessary because most gene trees had very low support values (and thus high estimation error); branch lengths estimated in coalescent units are directly impacted by observed gene tree discordance, and including gene trees with high estimation error (i.e. low support gene trees) inflates the amount of ILS.

In addition to the base model species tree, we created two other model species trees that had increased or decreased levels of ILS. To achieve this, we simply multiplied all branch length by 2 to reduce ILS, and divided them by 2 to increase ILS. Each of the resulting six model species trees with branch lengths in coalescent units were used to simulate gene trees using Dendropy (*58*) and based on the multi-species coalescent model (*20*).

Branch lengths on the simulated gene trees are expressed in coalescent units, and have to be converted into expected numbers of substitutions for simulating sequence alignments. To do this conversion, we used branch lengths observed from trees reconstructed from real data. For the avian data set, we used the gene trees reconstructed from the 190 longest introns. For the mammalian dataset, we used all 447 gene trees from (*6*). For branches leading to leaves, we used the species names to map the values from the reconstructed gene trees onto the simulated gene trees, randomly picking a branch length among all gene trees under consideration. For internal branches, we used a different approach. We ordered the internal branch lengths on the simulated trees and on the reconstructed trees separately, and matched branch lengths from the reconstructed trees with branch lengths from the simulated trees by their rank percentile. Thus we converted branch lengths on simulated gene trees such that their branch in each specific rank percentile had the same length as the reconstructed gene trees of the real dataset. This way both internal and external branches of the simulated gene trees had realistic branch lengths, as observed in the real data. Also, this produced model gene trees that are not ultrametric (i.e., do not exhibit the strong molecular clock).

For each of the resulting simulated gene trees, we simulated alignments under a GTR+G4 model using bppseqgen (*59*) based on parameters estimated by bppml (*59*) on the subset of avian genes that had all the taxa (1185 genes). The same GTR parameters were used for the mammalian dataset.

The parameters of bppseqgen for our simulations were:

- The substitution model parameters (GTR parameters):
  $a = 1.062409952497, b = 0.133307705766, c = 0.195517800882,$
  $d = 0.223514845018, e = 0.294405416545,$
  $\theta = 0.469075709819, \theta 1 = 0.558949940165, \theta 2 = 0.488093447144$
- The rate distribution parameters (Gamma parameters):
  $n = 4, \alpha = 0.370209777709$

For the avian simulation, we created conditions that reflect the four partitions of the biological data: exons-only, UCEs-only, introns-only, and long introns-only (restricted to loci with at least 10,000 sites) with respect to average bootstrap support of these partitions. To achieve these bootstrap support values, we simulated sequence alignments with 250bp, 500bp, 1000bp, and 1500bp, which resulted in average BS of 27%, 37%, 51%, and 60% – values that are very close (see Fig. S1) to those of the four partitions of the avian datasets (24%, 39%, 48%, and 59%). For the mammals dataset, we used sequence lengths of 500bp and 1000bp, which resulted in average BS values of 63% and 79%, effectively bracketing the average BS in the real dataset (71%). Note that to get shorter alignments, we simply trimmed our longest alignments (1500bp) by retaining the first 250bp, 500bp, or 1000bp sites and discarding the rest. Thus, when we change gene tree support by changing alignment length, each sequence alignment is simply a subset of the alignment used at higher support levels. Also note that the avian simulation represents a much harder condition than the mammalian simulation for two reasons: the avian gene trees have much higher estimation error than the mammalian simulation, and there is a higher amount of ILS (Table S1).

In addition to varying the amount of gene tree support and ILS levels, we created datasets with various number of genes. We simulated 20,000 gene trees for the avian dataset, and 4,000 genes for the mammalian dataset. We then sampled these genes (without replacement) to create model conditions with varying number of genes: 200 genes (20 replicates), 500 genes (20 replicates), 1000 genes (20 replicates), and 2000 genes (10 replicates) for the avian, and 200 genes (20 replicates), 400 genes (20 replicates), and 800 genes (20 replicates) for the mammals. Higher numbers of genes are explored for the avian experiment because it presents a more challenging model condition: even with 2000 genes, reconstructed species trees still have considerable error.

Finally, we built one replicate of a model condition with 14,350 genes for the avian simulation experiment in order to closely approximate the actual avian dataset in terms of the number of loci and average bootstrap support for estimated gene trees; thus 8250 genes are exon-like in terms of average bootstrap support, 2500 are intron-like, and 3600 are UCE-like. Similarly, we built a mixed model condition for mammals, where 200 genes of 63% support level and 200 genes of 79% support level were combined to get 400 genes of 71% average support, resembling the real dataset.

1.3 Evaluation procedure

We measure gene tree error for individual bootstrap replicates of gene trees using the missing branch rate (also called the False Negative rate), which is the percentage of the internal branches in the true tree that are missing in the estimated tree (i.e., we do not consider branches that have a leaf as one of their endpoints). We note that bootstrap replicate gene trees estimated using RAxML are always fully resolved, and hence the missing branch rate is identical to the standard Normalized Robinson-Foulds (RF) rate.

We also measure how well the entire distribution on gene trees is estimated. To compare gene tree distributions, we calculate how often each of the three possible topologies for every triplet of taxa appears in the set of true gene trees and the set of estimated bootstrap gene trees and supergene trees. Thus for every triplet, we get a distribution based on true gene trees and another one based on estimated gene trees, and we use the Kullback-Leibler (*60*) divergence statistic to measure how much the estimated distribution diverges from the distribution based on true gene trees.

We measure species tree topological error using both the missing branch and false positive rates. In the vast majority of the cases, the estimated trees are fully resolved, and so the missing branch (false negative) and false positive rates are equal. In a few replicates of the avian simulated dataset, the species trees were incompletely resolved (so instead of 45, only 43 or 44 branches were present in the estimated species tree), and in those cases false positive rates are slightly smaller than the missing branch rates. The general pattern of performance does not change whether error is measured by missing branch (false negative) or false positive rates.

Finally, we measured the error in estimation of coalescent unit branch lengths estimated by MP-EST. We computed the ratio of the estimated length to true length for those estimated branches that are also true species tree branches; since branches on MP-EST trees are in coalescent units, this evaluation also addresses how well the amount of ILS is estimated by the method.

1.4 Multi-locus bootstrapping procedure

We use the site-only multi-locus bootstrapping (*36*) for all summary methods, implemented as follows. For each gene or supergene, 200 replicates of bootstrapping is performed using RAxML. Then, 200 different inputs to the summary method are built, where each of these 200 inputs consists of the i^{th} bootstrap replicate across all genes, with $1 \leq i \leq 200$. Next, the summary method (MRP, MP-EST, or Greedy) is run on each of the 200 inputs, and 200 "bootstrapped" species tree replicates are obtained. A greedy consensus of these 200 bootstrap species tree replicates is built, and support values are drawn on this greedy consensus by counting occurrences of each bipartition in the 200 replicates.

1.5 Gene and supergene tree estimation

All supergene trees were estimated using RAxML with a procedure similar to unbinned gene trees (GTRGAMMA model, 200 bootstrap replicates). It is possible to use partitioning in the estimation of supergene trees, so that branch lengths and other model parameters can be estimated separately for each gene. However, due to computational concerns, we did not perform partitioning on any of the simulated datasets.

Supergene trees on the avian, yeast, vertebrates, and metazoa datasets were estimated using a partitioned analysis, assigning one partition per gene. This partitioning was required because these datasets included amino-acid sequences (which requires model selection). Model selection had already been performed on these datasets in their respective publications, and thus we had access to the selected model for each gene from these studies. The mammalian dataset was entirely composed of DNA sequences, and we used an unpartitioned GTRGAMMA RAxML analysis on this dataset.

1.6 Concatenation analyses

The concatenation analyses of the simulated datasets were performed using an unpartitioned RAxML GTRGAMMA maximum likelihood analysis with 20 independent runs with varying random seed numbers, but without bootstrapping. Concatenation analyses of the biological datasets were obtained from the relevant publications, with the exception of the mammalian dataset analysis on the reduced gene dataset, which we re-estimated using an unpartitioned RAxML GTRGAMMA maximum likelihood analysis.

1.7 Commands and version numbers

1.7.1 Estimating ML gene trees

We used RAxML version 7.3.5 (*33*) to build gene trees and perform bootstrapping.

Maximum likelihood trees:

```
raxmlHPC-SSE3 -m GTRGAMMA
-s [input_MRP_file] -n [a_name] -N 20
-p [random_seed_number]
```

Bootstrapping:

```
raxmlHPC-SSE3 -m GTRGAMMA
-s [input_MRP_file] -n [a_name] -N 200
-p [random_seed_number] -b [random_seed_number]
```

1.7.2 MP-EST

MP-EST version 1.0.3 was used in all runs. We used a custom shell script to run MP-EST 10 times with different random seed numbers and take the tree with the highest likelihood. For estimating branch length on a fixed topology (used in the simulation procedure) we used version 1.0.4 of MP-EST.

1.7.3 MRP

MRP data matrices are built using a custom Java program available at `https://github.com/smirarab/mrpmatrix`. The following command was used to create the MRP matrix.

```
 java -jar mrp.jar [input_file] [output_file] NEXUS
```

The default heuristic in PAUP* (v. 4. 0b10) (*61*) was used for solving the parsimony problem. This heuristic operates by first generating an initial tree through random sequence addition and then using Tree Bisection and Reconnection (TBR) moves to reach a local optimum. 1000 iterations are used, and the most parsimonious tree is returned. When multiple trees have the same maximum parsimony score, the greedy consensus of those trees is returned. The following shows the PAUP* commands used.

```
 begin paup;
set criterion=parsimony maxtrees=1000
increase=no;
hsearch start=stepwise addseq=random
nreps=100 swap=tbr;
filter best=yes;
savetrees file = <treeFile> replace=yes
format=altnex;
contree all/ strict=yes
treefile = <strictConsensusTreeFile>
replace=yes;
tcontree all/ majrule=yes strict=no
treefile = <majorityConsensusTreeFile>
replace=yes;
contree all/ majrule=yes strict=no
le50=yes
treefile = <greedyConsensusTreeFile>
replace=yes;
log stop;
quit; end;
```

1.7.4 Greedy

We use Dendropy version 3.12.0 (*58*) to compute the greedy consensus tree.

1.7.5 Binning

A set of custom shell and Python scripts are used to build the binning pipeline. The entire pipeline is publicly available at www.cs.utexas.edu/users/phylo/datasets/binning/. Each step is described below.

Finding compatibility between pairs of gene trees was performed using custom JAVA code.

Once compatibility is measured, the incompatibility graph is built using custom Python code, and the code for finding balanced vertex coloring is invoked. For balanced vertex coloring, we modified a publicly available implementation of the vertex coloring algorithm developed by Shalin Shah (http://shah.freeshell.org/graphcoloring/). The vertex coloring defines the bins. We then use a custom perl script to build concatenated alignments for each supergene. Finally, we use RAxML to estimate supergene trees, using the same commands used for estimating gene trees.

2 Supplementary Text

In this section, we provide additional explanation of results obtained both on biological and simulated datasets. Some of these results are presented only in supplementary figures due to space constraints.

2.1 Results from simulation studies

We compare the accuracy of species trees estimated using concatenation using maximum likelihood (computed using RAxML (*33*)), and three summary methods (with and without binning). The summary methods we explore are the greedy consensus (also known as the extended majority consensus) (*40*), Matrix Representation with Parsimony (MRP) (*35*), and MP-EST (*13*).

2.1.1 Performance of MRP and Greedy

For most avian model conditions, Greedy had the lowest accuracy, MP-EST had the best accuracy, and MRP was in between Greedy and MP-EST in terms of accuracy (see Figure S6). On the mammalian dataset all error rates are lower, and the relative performance of MRP, Greedy, and MP-EST is more mixed; each method is sometimes the best, but generally the differences are not substantial.

Binning generally improves both Greedy and MRP, similar to the pattern observed for MP-EST. However, there are three cases where binning reduces the accuracy of MRP by a small margin: 400 and 800 mammalian genes with 79% support and 1000 long intron-like avian genes.

In a few rare cases on the avian datasets, the greedy consensus trees produced by the multi-locus bootstrapping procedure (see Section 1.4) had one or two edges missing, and so the trees had small polytomies. In such cases, the missing branch (false negative) rate and false positive rates will be slightly different. For completeness, Figure S7 shows false positive rates for all the cases shown in Figure 2 of the main paper; note that these false positive results are for the most part indistinguishable from missing branch rate results, and that relative performance between methods does not change. Note also that on the mammalian dataset, multi-locus bootstrapping never produced polytomies, and hence false positive rates are identical to false negative rates.

2.1.2 Gene tree estimation accuracy

Statistical binning allows gene trees to be re-estimated, because after the bins are computed, each gene is associated with a supergene. Therefore, the supergene tree estimated on the supergene alignment can be used as a revised estimate of the gene tree for that gene.

We explored the impact of binning on the estimate of the rooted gene tree distribution; see Figures S4 and S5. We represent each estimated gene tree distribution (defined either by the set of true gene trees or the set of estimated gene trees) by the frequency of each of the three possible alternative topologies for all the $\binom{n}{3}$ triplets of taxa, where n is the number of taxa. As a result, for each replicate of each dataset, we have $\binom{n}{3}$ true triplet distributions, and a binned and an unbinned estimated triplet distribution. We calculated Kullback-Leibler (KL) (*60*) divergence of each of these $\binom{n}{3}$ distributions that are based on estimated gene trees or supergene trees from the corresponding distributions estimated from true gene trees. Note that we do not have the true distribution on triplets, but rather use the set of true gene trees to estimate the true distribution on triplets. We thus obtain $\binom{n}{3}$ KL divergence values, which we plot as boxplots.

The improvement in the gene tree distribution estimation is largest when ILS is low and the sequence lengths are short, but even present when the sequence lengths are longer and ILS is high (both of which reduce the bin sizes), as shown in Table S1.

Figures S2 and S3 show Robinson-Foulds (RF) distances between the true gene trees and estimated gene trees, considering both initial estimated gene trees and the result of using the estimated supergene trees produced by binning. Since the input to the summary methods that estimate species trees are the bootstrap replicates of each gene tree, we calculate the RF distance to all the 200 estimated bootstrap replicate gene trees. As Figures S2 and S3 show, these re-estimated gene trees are generally more accurate than the original estimates.

2.1.3 Concatenation vs. coalescent-based estimation

On the avian datasets, MP-EST analyses on true gene trees had uniformly better accuracy than concatenation (in many cases dramatically better accuracy). The only cases where Greedy analyses on true gene trees were more accurate than concatenation were with the smallest number of genes (200) we tested.

However, performance on estimated gene trees revealed important differences between summary methods and concatenation. On all the avian datasets we explored, Greedy produced less accurate estimations than concatenation, with differences that ranged from 4% (200 UCE-like genes) to 23% (the mixed dataset). While binning improved Greedy, in most cases binned Greedy was still less accurate than concatenation. Thus, both binned and unbinned Greedy were clearly inferior to concatenation on the avian datasets.

Unbinned MP-EST was often less accurate than concatenation on the avian datasets, and the only cases where unbinned MP-EST was more accurate than concatenation were with large numbers of relatively well estimated gene trees (e.g., 1000 long intron-like genes). In contrast, binned MP-EST typically either matched or improved on the accuracy of concatenation. For example, on 1000 exon-like genes, binned MP-EST had about 10% less error than concatenation. On the mixed dataset, concatenation and binned MP-EST had the same error rate (7%), while all the other methods (unbinned MP-EST, and both binned and unbinned Greedy) had at least 11% error.

On the mammalian datasets, the relative performance between concatenation and summary methods was closer. All methods had good accuracy (no error rates above 8%), which reduced the differences between methods. However, some differences can still be noted. Although summary methods applied to true gene trees produced results that were more accurate than concatenation, the performance varied on estimated gene trees. On the gene trees with lower accuracy (reflected in 63% average bootstrap support), species trees estimated using unbinned Greedy or unbinned MP-EST were less accurate than concatenation for almost all cases; the only exception was unbinned Greedy on 200 genes, where it matched the accuracy of concatenation. Binned summary methods had better accuracy than concatenation in all cases (however, differences were small with 400 genes). On the higher BS gene trees, summary methods tended to either match (800 genes) or improve on concatenation (200 and 400 genes); however, the differences were small (at most 2%). On the mixed dataset with 400 genes, only unbinned MP-EST was less accurate than concatenation, and binned MP-EST had the best accuracy; however, the differences on these mixed mammalian datasets was again small, at most 2%.

Thus, on the mammalian datasets, summary methods - especially the binned versions - were typically more accurate than concatenation, but with small differences, while on the avian datasets only binned summary methods, especially MP-EST and MRP, typically improved or matched concatenation. The impact of statistical binning on the relative performance of concatenation and MP-EST was most significant when there was a large number of genes, except when the gene trees are very accurately estimated. As an example, on 2000 UCE-like avian genes, unbinned MP-EST had 16.4% error, concatenation had 8.4% error, but binned MP-EST had only 6.7% error. Similarly, on 2000 Intron-like avian genes, unbinned MP-EST had 8.2% error, and concatenation had 8.4% error, but binned MP-EST had 3.3% error.

2.2 Results on biological datasets

Additional results on the biological datasets are presented in Figures S12 to S21, and further explained here. Figure S12 shows the size of bins for various biological datasets. Figures S13 and S14 show the impact of binning on gene trees, and the remaining figures presents additional results corresponding to individual biological datasets.

Impact of binning on biological gene trees: Figure S13 shows the amount of discordance between gene trees of various biological datasets, with and without binning. Figure S14 shows bootstrap support in the estimated gene trees for the mammalian, metazoan, vertebrate, and yeast biological datasets; note that without binning, the metazoan dataset gene trees have extremely low bootstrap support, but other datasets also have many genes with low bootstrap support. For example, at least 40% of branches in the majority of genes in all datasets have less than 75% support (to see this note that median bars do not exceed the 60% mark for all datasets in Figure S14B). The average BS was 71% for the mammalian dataset, 49% for the metazoa, 76% for the vertebrates, 72% for yeasts, and 32% for the the avian dataset. Binning improves the bootstrap support of the estimated gene trees - with large improvements for mammals, metazoa, and avian datasets, and small improvements for vertebrates and yeast. The difference in impact is due to smaller bins for the vertebrate and yeast datasets, resulting from the general better resolution in their unbinned gene trees. Thus, when bins are larger, the supergene trees are based on longer sequence alignments, and have higher support.

Put together, these two figures show that although there are high levels of discordance among estimated gene trees, some of the discordance is likely due to estimation error reflected in the low average BS values, rather than being purely due to biological processes such as ILS. These observations are consistent with ones made on simulated datasets.

2.2.1 Avian dataset analyses

Evidence of ILS: Evidence for ILS in the avian dataset is extensively reported in (*31*). Here we present some statistics, but the reader is referred to that publication for a more in-depth discussion of ILS in the avian dataset. The average topological distance between estimated gene trees and the TENT (the RAxML tree on the concatenated alignment with the full set of approximately 14K genes) was very high (74%). However, most markers (exons, introns, and UCEs) had low phylogenetic signal, with the result that the average bootstrap support (BS) for the estimated gene trees was very low, only 32%. Among the 14,446 gene trees, introns had the highest BS values (48%), and also had a somewhat lower distance to the concatenation tree (63%). Thus, the introns had relatively low average BS values, but the other partitions had even lower support (exons had 24% BS and UCEs had 39% BS). Therefore, the large topological distances between estimated trees is to some extent a result of poor phylogenetic signal in the gene sequences.

However, as we will show, substantial evidence of strongly supported gene tree conflict remained even after taking these low bootstrap support values into account. First, as can be see in Figure S15, many of the branches in the TENT are rejected by a large the number of intron gene trees with high support (at least 75%); furthermore, there are many short branches adjacent to each other in the tree, as expected in a rapid radiation scenario. Also, this is a condition that leads to anamalous gene trees (*9*) (where the most probable gene tree topology is not identical to the species tree). Similarly, comparing gene trees to each other revealed substantial levels of discordance. The second piece of evidence for strongly supported gene tree conflict is that, on average, two estimated intron gene trees differed in 1.3 strongly supported edges (at least 75% support). Thus, a very high level of discordance is observed in the avian dataset, some of which is clearly due to lack of support. However, a lot of discordance is observed even among highly supported branches, providing evidence of real gene tree discord.

Species Trees The results on the avian dataset are reported in depth in (*31*). Here we briefly discuss the effects of binning on the avian biological dataset (see Figure S16). The unbinned MP-EST tree on the full set of approximately 14K genes (i.e., the same input as for the TENT) had low support, and contradicted most other rigorous analyses reported in that paper. More specifically, four novel clades – Australaves, Cursores, Otidimorphae, and Columbea, see Figure S16 for the definition of these clades – consistently showed up in concatenation analyses that included introns and UCEs, and also unbinned MP-EST analyses restricted to non-coding data, but were not recovered by the unbinned MP-EST analysis on the TENT matrix. However, the input gene trees, especially the exons, had very low support (average bootstrap support was 25% for exons, 35% for UCEs, and 48% for introns), and thus having low support edges in the unbinned MP-EST species tree is not surprising. The binned MP-EST tree of the TENT matrix was well resolved and recovered the four clades that were missing in the unbinned MP-EST tree on the TENT matrix. The binned MP-EST tree on the TENT matrix was not identical to the TENT, but was more congruent with it (Fig. S17). The binned MP-EST tree on the TENT matrix and the TENT tree are the two major hypotheses in (*31*), and are discussed in detail in that paper.

Interestingly, both unbinned and binned MP-EST analyses of the intron dataset recover Australaves, Cursores, Otidimorphae, and Columbea, just like the binned MP-EST on all 14K genes and the concatenation results on datasets that included non-coding sequence. Thus, these intron-only MP-EST trees are more congruent with other reliable analyses (but see our companion paper for an in-depth discussion of the remaining incongruence between these analyses (*31*)). This similarity is likely because intron gene trees have better support than other other partitions, and (as shown in our simulation study) when gene trees have high support, even unbinned MP-EST can have high accuracy.

2.2.2 Mammalian dataset analyses

We report results for a re-analysis of the dataset studied in (*6*). We filtered 21 problematic genes that we identified as having mislabeled species names (this was subsequently confirmed by the authors), plus 2 genes that were clear outliers (see Figure S18). The filtering of these two outlier genes was required for the binning procedure, since they were placed in bins by themselves (each in a bin of size one), whereas the average bin size was 6.5. Including the outlier genes in the binned MP-EST analysis would have produced a very distorted distribution on triplet gene trees, and reduced the accuracy of the MP-EST analysis, since it depends on accurate distributions on triplet gene trees.

Figure S19 shows analyses of the Mammalian dataset after removing these 23 problematic genes, and compared to the original analysis in (*6*). Binned and unbinned MP-EST trees are topologically identical except for the position of tree shrews (*Tupaia Belangeri*). Bootstrap support is generally higher in the binned MP-EST tree, except that the bootstrap support for bats (*Myotis lucifugus* and *Pteropus vampyrus*) is lower in the binned MP-EST tree (82.5%) than in the unbinned MP-EST tree (94.5%). The concatenated analysis without the 23 problematic genes is topologically identical to that reported by (*6*), with slight differences in bootstrap support. Concatenation and binned MP-EST trees are identical topologically except for the location of bats (*Myotis lucifugus* and *Pteropus vampyrus*).

2.2.3 Metazoan dataset analyses

The most important distinction between the binned and unbinned MP-EST trees is among Chordates, where the unbinned analysis puts Cephalochordates (represented by *B. Floridae*) as sister to vertebrates (Craniates), but the binned analysis puts Urochordates (represented by *C. intestinalis*) as sister to vertebrates. The relationship retrieved by the binned MP-EST has overwhelming support in the recent literature based on molecular studies (*44–46*), and is likely correct. However, the assumption before these recent studies was the Cephalochordates/vertebrates relationship (*43*).

Sister to Bilateria: In both the binned and unbinned MP-EST trees, *N. vectensis* (representing Cnidaria) is grouped with *T. adhaerens* (representing Placozoa), and these two are sister to Bilateria. This relationship, which contradicts the monophyly of Eumetazoa, has some support in the literature (*62*), but the majority of recent molecular studies are congruent with the relationship recovered in the concatenation tree, where *N. vectensis* is sister to Bilateria (*63,64*).

Protostomia: There are also some differences in the binned and unbinned MP-EST trees with respect to Protostomia, but these are hard to interpret because some relationships among major lineages of Protostomia are not well established (*65*). Concatenation (see Fig. S20), binned and unbinned MP-EST analyses each results in a different resolution for Protostomia, and no topology is identical to some of the newer molecular studies (*63, 65*). This is likely due to the poor taxon sampling of this dataset (only 20 metazoan taxa).

In all trees, Annelid (represented by *H. Robusta*) and *Mollusca* (represented by *L. gigantea*) are sisters with full support, as expected. However, Nematoda (represented by *C. elegans*), and Platyhelminthes (represented by *S. mansoni*) are put in different places. The likely correct relationship is that Nematoda should be sister to Arthropoda, and Platyhelminthes sister to Mollusca/Annelid (*65*). The unbinned MP-EST analysis puts Platyhelminthes as sister to Mollusca/Annelid with 70% support, but fails to put Nematoda as sister to Arthropoda. Binned MP-EST recovers neither relationship, but is in fact essentially unresolved with regard to the relationship between Mollusca/Annelid, Platyhelminthes, and Arthropoda (only 32% support for an Arthropoda/Mollusca/Annelid clade, and 54% for Nematoda/Platyhelminthes). Concatenation puts Nematoda as sister to Platyhelminthes.

Among Arthropoda, the binned and unbinned MP-EST trees differ in the position of Hymenoptera (represented by *A. mellifera*), where the binned MP-EST tree puts them as sister to other Holometabola, but the unbinned MP-EST tree puts them as sister to Coleoptera (represented by *T. castaneum*). While the exact position of Holometabola continues to be debated, recent molecular analyses are consistent with the position in the binned MP-EST tree (*66*).

2.2.4 Vertebrate dataset analyses

The binned and unbinned MP-EST trees are topologically identical to each other and to the concatenation tree (Fig. S21); the only difference is the bootstrap support for the clade containing horse (*E. caballus*) and dog (*C. familiaris*). The unbinned analysis has higher support (97%) for this clade, and the binned analysis has lower support (83%). All other branches have 100% support in both analyses. Whether horses (and more generally Perissodactyla) are closer to dogs (more generally Carnivora) or cows (more generally Cetartiodactyla) is an open question; see (*67*) for a comprehensive summary.

2.2.5 Yeast dataset analyses

The binned and unbinned MP-EST topologies were identical, and both had 100% support for all but one branch (Fig. S22). Both trees were also identical to the concatenation tree reported in (*32*) in all branches, except for the single branch that had less than 100% support. This particular branch unites *C. lusitaniae* with the *C. guiliermondii*/*D. hansenii* clade. While the exact position of *C. lusitaniae* is not known, the relationship recovered in the MP-EST trees is closer to current belief about yeast evolution (*68*).

2.3 Computational resources and issues

2.3.1 Computational requirements of statistical binning

A statistical binning analysis of a single dataset (i.e. one replicate of a particular model condition) involves the following steps:

Step 1: estimate unbinned gene trees with bootstrapping,

Step 2: compare all pairs of genes for combinability,

Step 3: run our heuristic for balanced vertex coloring and create supergene alignments,

Step 4: estimate ML (maximum likelihood) supergene trees with bootstrapping, and

Step 5: build supergene trees using three summary methods: MP-EST, MRP, and Greedy.

Of these, the most computationally demanding steps are Steps 1, 4, and 5. Steps 1 and 4 involve running RAxML with bootstrapping in order to estimate ML gene trees and supergene trees with branch support. Step 5 is expensive only because of the MP-EST analysis (Greedy and MRP are extremely fast). All other steps, including running the heuristic for balanced vertex coloring, are relatively much faster than Steps 1, 4, and 5.

Step 1: Estimating a single gene tree with bootstrapping typically took between 30 to 120 minutes for avian genes (depending on alignment length), and between 15 to 30 minutes for the mammalian dataset. Thus a single simulated dataset of 2000 avian genes would require between 40 to 160 days of serial computation. The simulated mammalian datasets were faster because they had fewer genes (at most 800), and so could finish in 200 to 400 hours (8 to 16 days of serial computation). The gene tree estimations can obviously be done in parallel, and parallelization is trivial since the calculations are independent of each other. We used the Texas Advanced Computing Center (TACC) to run these analyses in parallel, using between 100 to 800 CPUs simultaneously.

Step 2: Calculating combinability was fast, taking, for example, less than 3 seconds to compare a single gene against all other genes for the 2000 avian genes model condition. Thus all pairwise comparisons with 2000 genes can be done in 90 minutes if done in serial; however, once again it is trivial to run these comparisons in parallel, as we did. We used a heterogeneous Condor cluster available to us from the Department of Computer Science at the University of Texas at Austin for running the combinability tests.

Step 3: Once combinability is determined, computing the graph and running the balanced vertex coloring heuristic is extremely fast, taking typically only a few seconds, and never more than a few minutes.

Step 4: Estimating supergene trees with bootstrapping can also be time consuming, but depends on the model condition. Where gene tree discordance is high and gene trees are very well-resolved (for example, long intron-like avian datasets), many genes form singleton bins, eliminating the need for re-estimation of supergene trees. When larger bins are formed, estimating gene trees for each supergene takes longer than a single gene, however, we have fewer supergenes than genes. Overall, when most genes are singleton, only a very small extra time is required to get supergene trees; however, when most bins have at least two genes, the time to compute supergene trees is comparable to the time required for estimating the original unbinned gene trees. For example, estimating all supergene trees for a single dataset of 2000 avian UCE-like genes took about 40 days of serial computation, or half a day with 80 parallel jobs.

Step 5: Running Greedy or MRP is fast, taking seconds for Greedy and minutes for MRP. Greedy is a low degree polynomial time algorithm, and very fast in practice. MRP is based on a heuristic for the NP-hard Maximum Parsimony problem, but again very fast in practice.

However, MP-EST, especially on the avian dataset, was very slow. MP-EST is also a heuristic for maximum pseudo-likelihood (here under the coalescent model), and the running time is strongly impacted by the number of taxa and the optimization landscape (and not particularly impacted by the number of genes). The running time for a single run of MP-EST on an avian dataset was not affected much by the number of genes, therefore, and took between half an hour and an hour per dataset. However, in our protocol, we use multiple runs of MP-EST to estimate a species tree on a single dataset. Specifically, we run MP-EST 10 times and take the tree with the best ML score that is found. Also, the multi-locus bootstrapping procedure requires running MP-EST on 200 distinct inputs. Thus, an MP-EST analysis of a single avian dataset consists of 2000 runs, each of which typically takes between half an hour to one hour to complete, which results in a total running time of 40-80 days if run sequentially (MP-EST runs were somewhat faster for the reduced ILS case, taking typically 30-50% less time, but was not largely impacted by increasing ILS, or changing gene tree support). MP-EST on the mammalian dataset was close to 5-10 times faster than avian (reflecting mainly the reduced number of taxa, but also the reduced gene tree conflict), requiring between 200 to 400 hours of serial computation for each dataset. Once again, multi-locus bootstrapping is trivially parallel. We used 200 parallel jobs run on our Condor cluster, and were able to finish each mammalian MP-EST analysis in 1-2 hours, and each avian analysis in 5-10 hours.

In summary, for a single 48-species avian dataset with 2000 genes of UCE-like support, a total of about 180 days of serial computation is needed for a binned MP-EST analysis: 80 days of serial computation time for the initial gene tree estimation, another 40 days for supergene tree estimation, and about 60 days for running MP-EST; all the other steps combined finished within 2 hours. Thus, in total about 180 days of serial CPU time for the binned MP-EST analysis. Using a cluster that can run 18 jobs in parallel, this can be done in less than 10 days. Running 60 jobs at a time would allow this to finish in about three days. The unbinned MP-EST analysis would take 140 days of serial CPU time, and could finish somewhat faster. Therefore, with moderate computational resources, binned MP-EST analyses with large numbers of genes and moderate numbers of taxa can be performed in reasonable timeframes.

2.3.2 Overall running time for this study

While any single analysis can be performed in a reasonable time with moderate amount of parallelism, this study involved tens of model conditions, and for each model condition we have looked at 20 replicates (10 replicates for the model conditions with 2000 genes). Thus our total computational time was extremely large: we estimate that we used more than 1,000,000 hours (or more than 100 years) of CPU time overall. Running all these analyses was doable only because of the exceptional computational resources we had access to at TACC supercomputers and the University of Texas Condor cluster.

3 Figs. S1 to S22

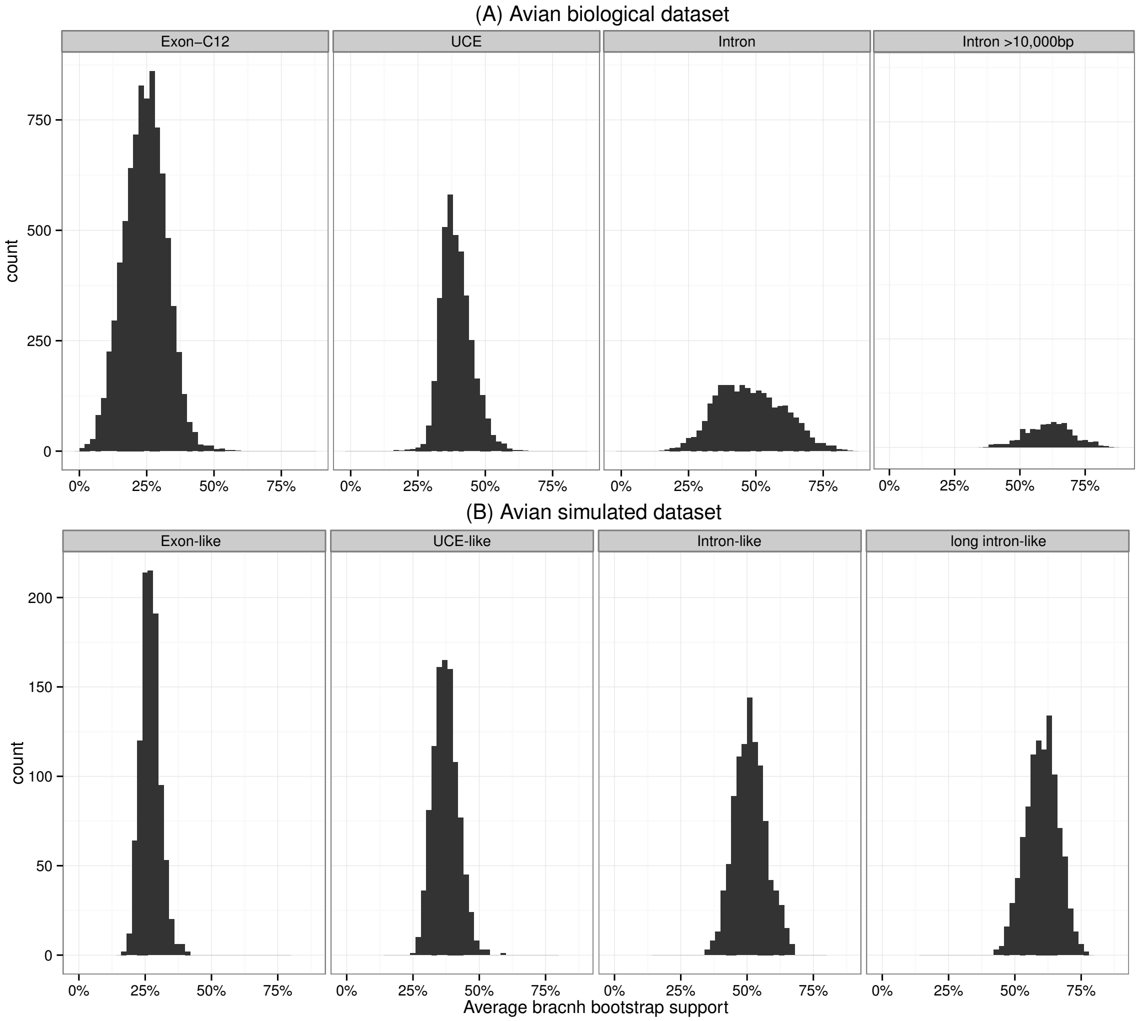


Figure S1: **Gene tree bootstrap support histograms for the avian simulated and biological datasets.** Histograms show the distribution of average bootstrap branch support across (A) three partitions of the avian dataset with a total of 14,446 loci (*31*), and (B) 1000 genes from each of the four simulated "support" model conditions for the avian dataset. Note the extremely low support of most loci in the avian dataset.

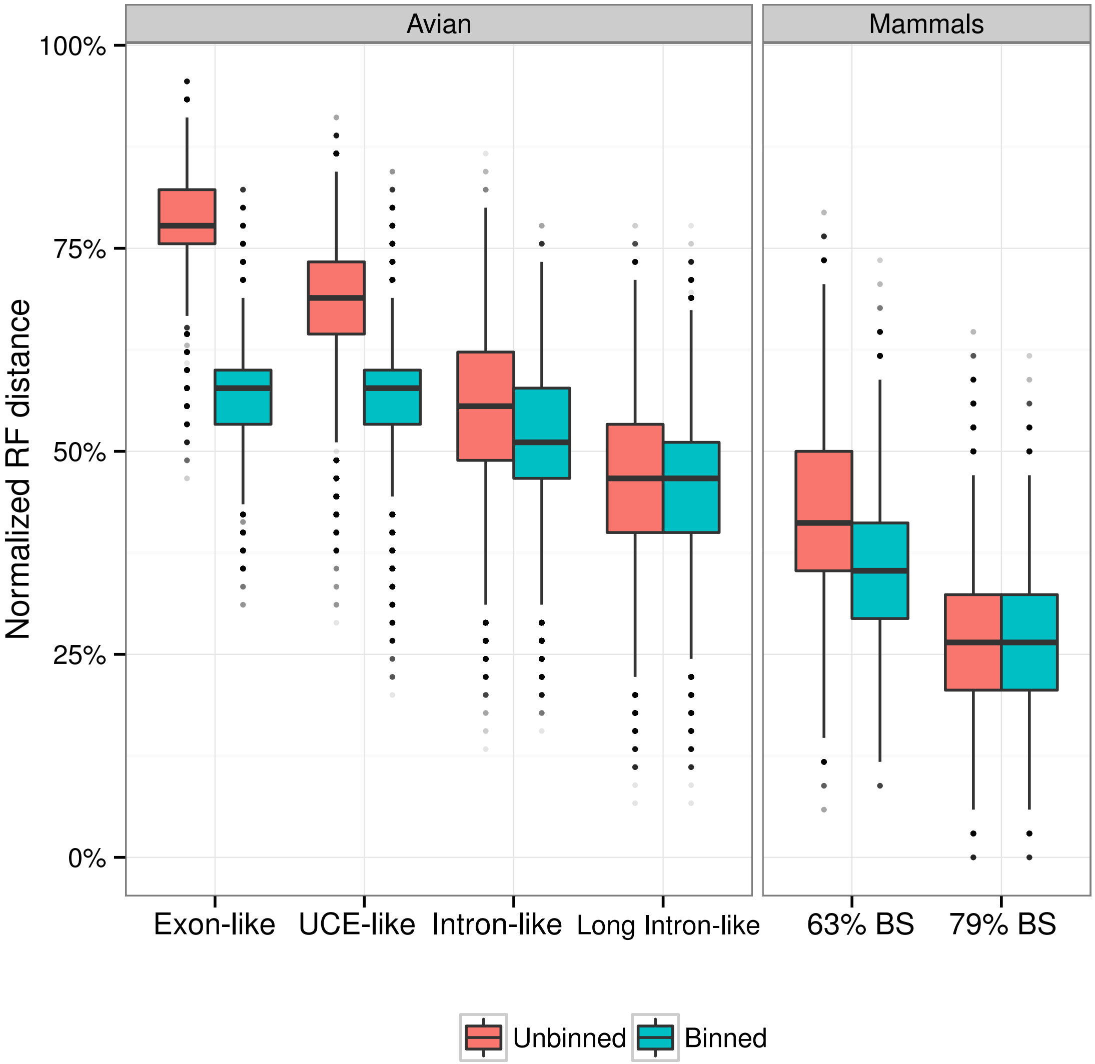


Figure S2: **Gene tree estimation error for various levels of gene tree support for the simulated 1X avian and 1X mammals datasets.** Results are shown for fixed number of genes (1000 for avian and 200 for mammals) and 1X ILS levels. We show the distribution of RF distances between true gene trees and all 200 bootstrap replicates of each estimated gene tree. Each bootstrap replicate of each supergene tree is compared separately against *each* true gene tree corresponding to genes put on that bin. Thus for both binned and unbinned gene trees, boxplots are over $200,000$ data points for the avian and $40,000$ data points for the mammalian datasets.

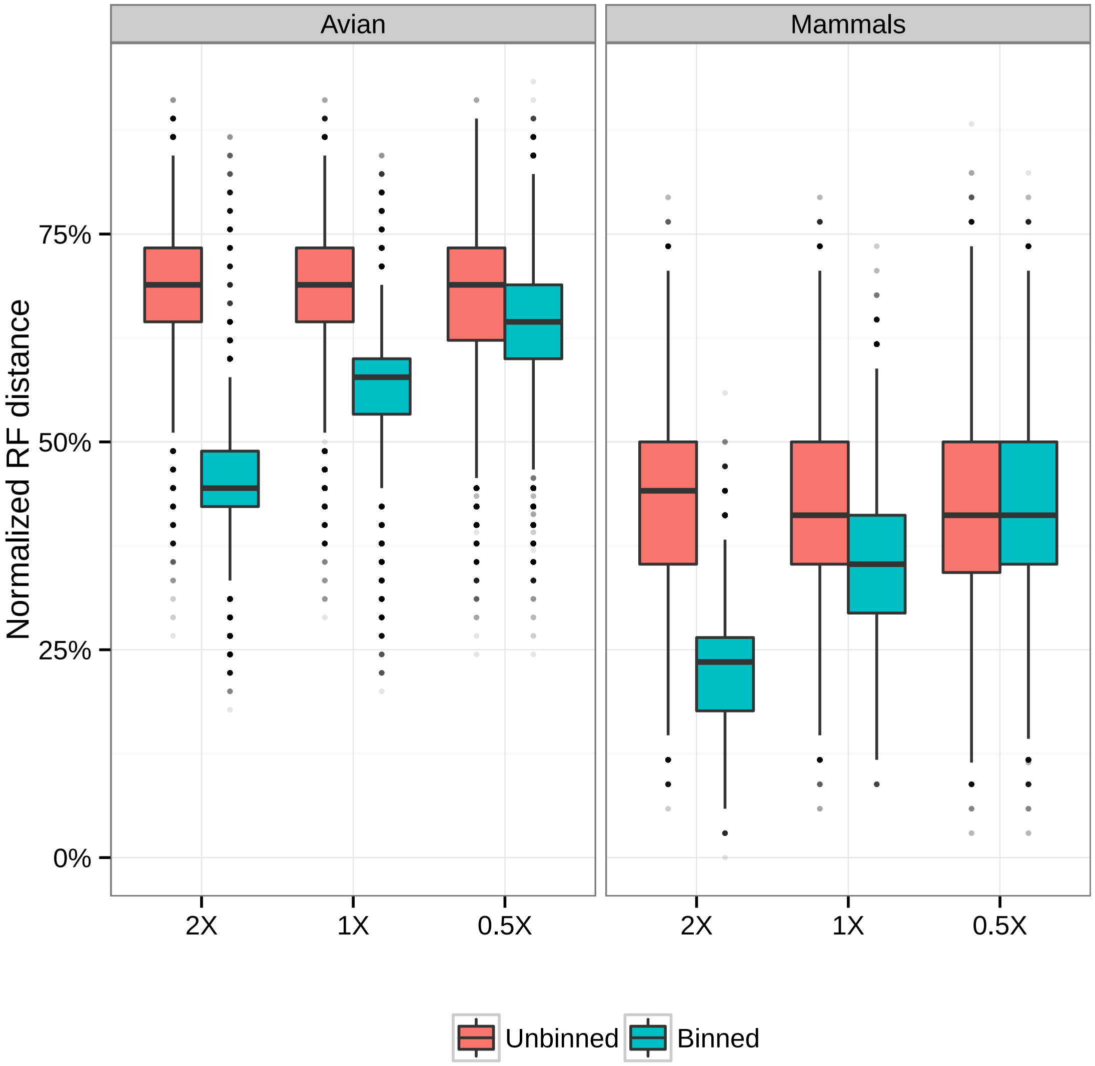


Figure S3: **Gene tree estimation error (computed using RF distances) as a function of ILS level.** We show results for 200 gene trees with 63% BS (moderate phylogenetic signal) for the mammalian simulated datasets, and for 1000 gene trees with UCE-like phylogenetic signal for the avian simulated datasets. The distribution of RF distances between true gene trees and all 200 bootstrap replicates of each estimated gene tree is shown. We show results for gene trees estimated with and without binning. Each bootstrap replicate of each supergene tree is compared separately against *each* true gene tree corresponding to genes put on that bin. Thus for both binned and unbinned gene trees, boxplots are over $200,000$ data points for the avian and $40,000$ data points for the mammalian datasets.

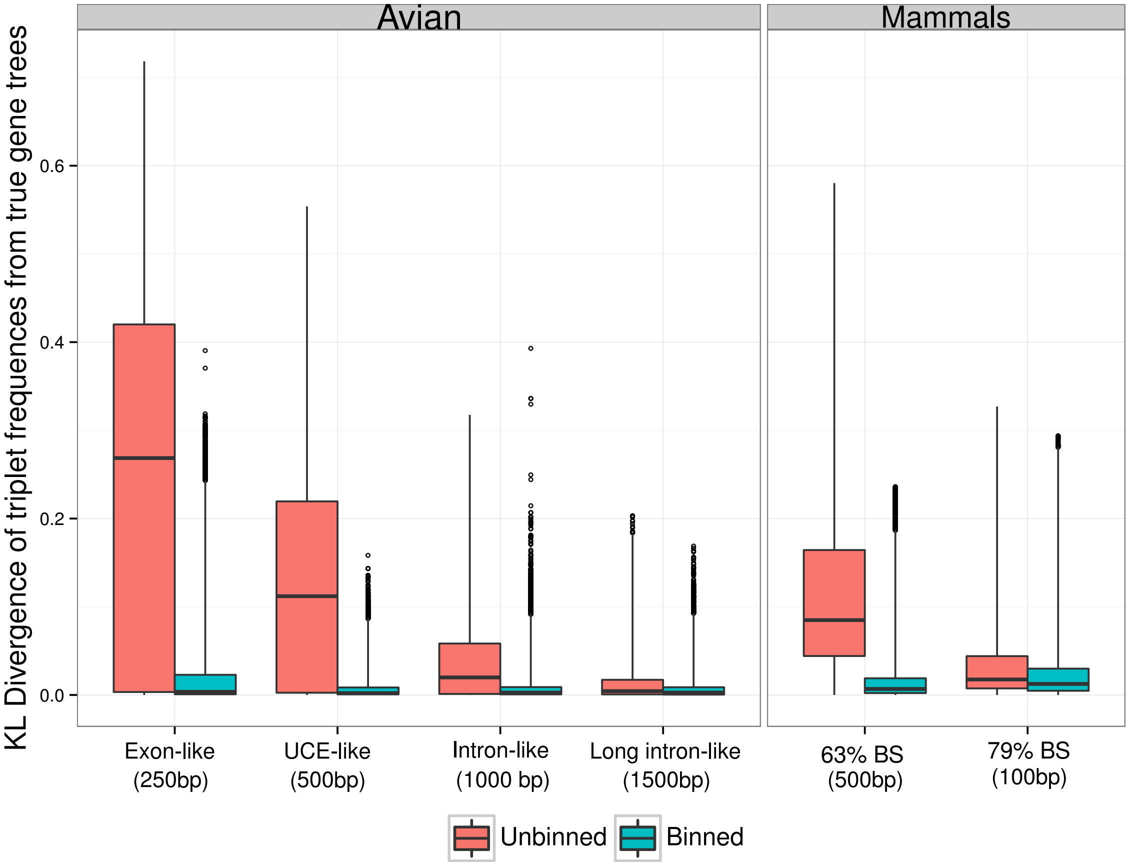


Figure S4: **Divergence of estimated gene trees triplet distributions from true gene tree distributions for simulated avian and mammals datasets with varying levels of gene tree support.** The number of genes is fixed to 1000 for avian and to 200 for mammals, and results are shown only for 1X ILS level. True triplet frequencies are estimated based on true gene trees for each of the $\binom{n}{3}$ possible triplets, where n is the number of species. Similarly, triplet frequencies are calculated for estimated gene trees using all 200 bootstrap replicates of all gene/supergene trees. For each of the $\binom{n}{3}$ triplets, the Kullback-Leibler (KL) divergence of the estimated triplet distribution from the distribution estimated using true gene trees is calculated. The boxplots show the distribution of these $\binom{n}{3}$ KL divergence measures, but results for the first 10 replicates of each dataset are aggregated into one boxplot (so each boxplot is over $10\binom{n}{3}$ KL measures, where $n = 48$ for avian and $n = 37$ for mammals). The whiskers are extended to 10 times the Inter Quartile Range range from each side, instead of the 1.5 times used by default in boxplots. This was necessary because these distributions are heavy-tailed and without extending the whiskers many points would be unjustifiably marked as outliers (*69*).

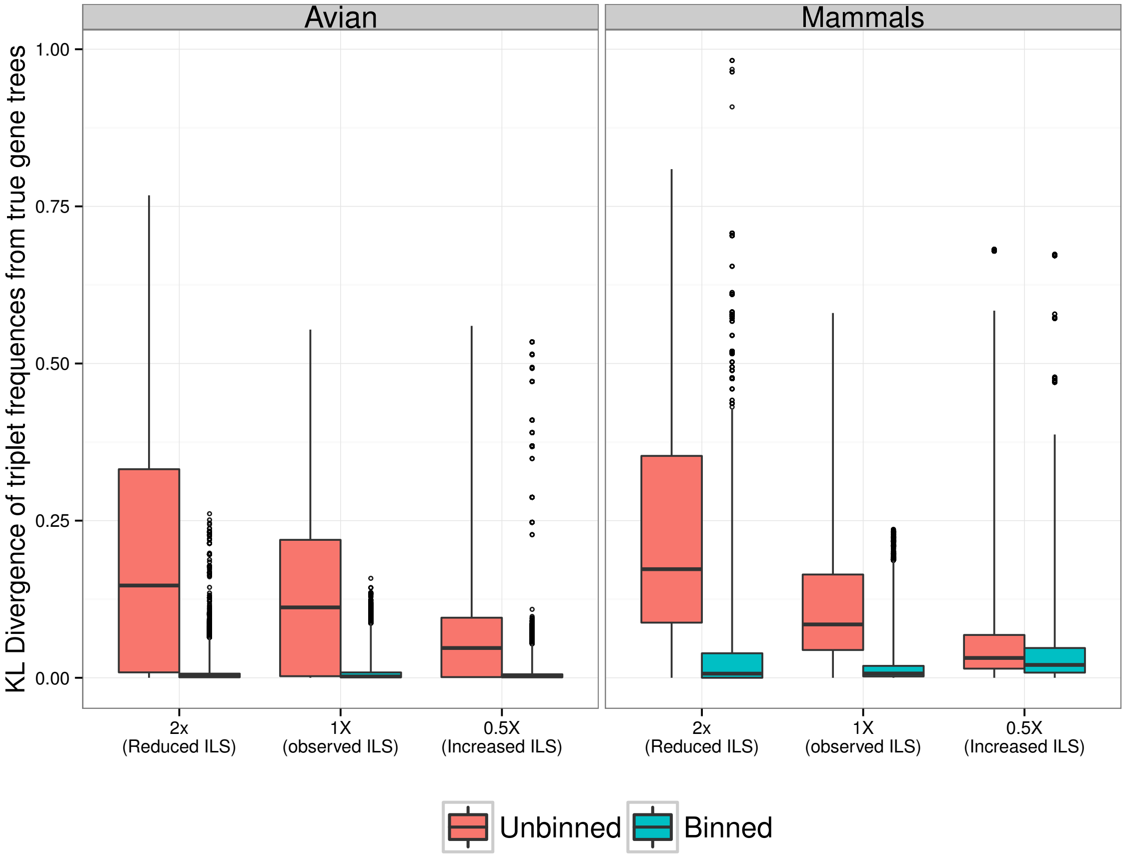


Figure S5: **Divergence of estimated gene tree triplet distributions from true gene tree distributions for simulated avian and mammals datasets with varying levels of ILS.** The number of genes is fixed to 1000 for avian and to 200 for mammals, and gene tree support is fixed to UCE-like for the avian dataset and to moderate support (63% BS) for the mammalian dataset. For each of the $\binom{n}{3}$ triplets of species, the true triplet frequency is estimated using true gene trees. Triplet frequencies are calculated for estimated gene trees using all 200 bootstrap replicates of all gene/supergene trees. For each of the $\binom{n}{3}$ triplets, the Kullback-Leibler (KL) divergence of the estimated triplet distribution from the distribution based on true gene trees is calculated. The boxplots show the distribution of these $\binom{n}{3}$ KL divergence measures, but results for the first 10 replicates of each dataset are aggregated into one boxplot (so each boxplot is over $10\binom{n}{3}$ KL measures, where $n = 48$ for avian and $n = 37$ for mammals). The whiskers are extended to 10 times the Inter Quartile Range range from each side, instead of the 1.5 times used by default in boxplots. This was necessary because these distributions are heavy-tailed and without extending the whiskers many points would be unjustifiably marked as outliers (*69*).

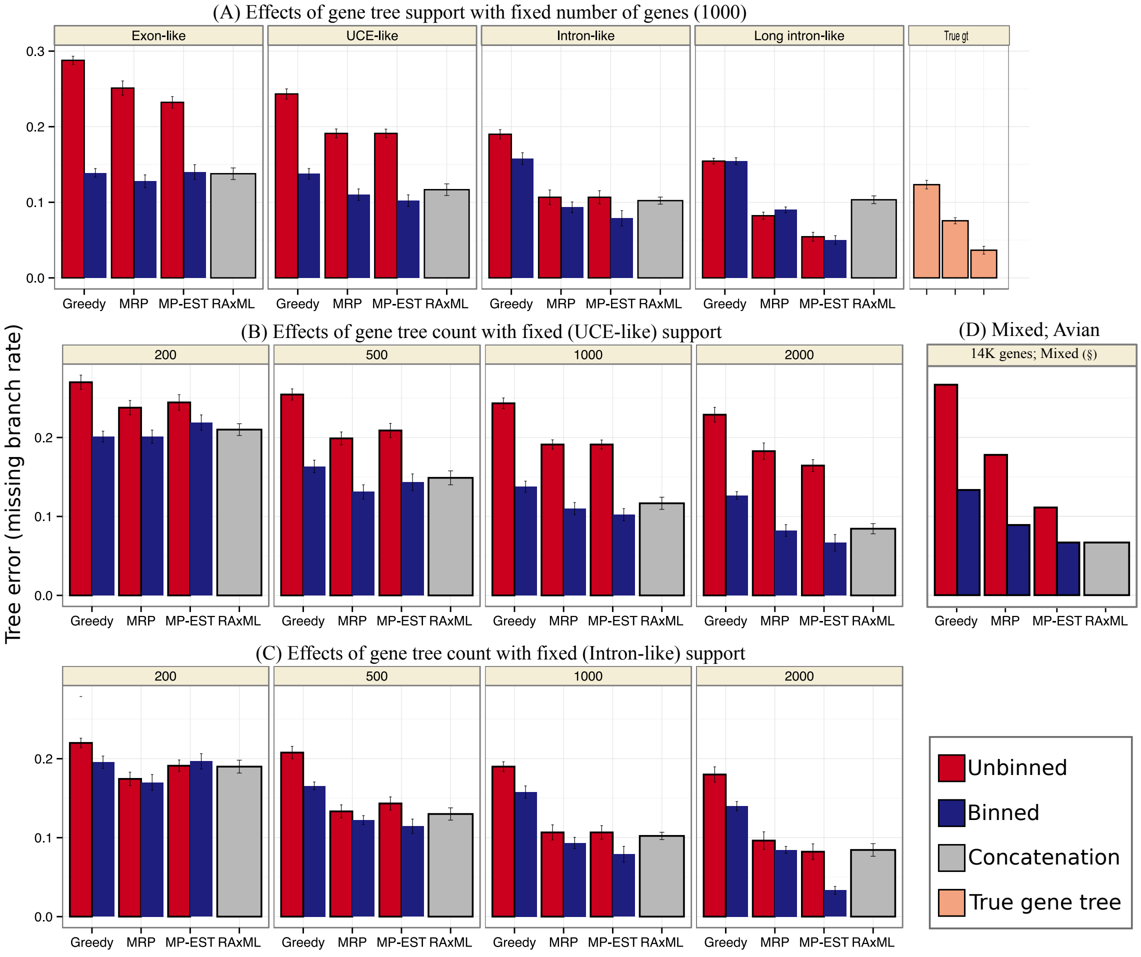


Figure S6: **Simulation results including MRP, Greedy, MP-EST, and concatenation using maximum likelihood on avian datasets.** Bars show average missing branch rates and error-bars show standard error. Results are over 20 replicates everywhere except 2000 genes model conditions, which is based on 10 replicates, and the mixed model condition, which is based on only 1 replicate. This figure shows results similar to those shown in Figure 2 of the main paper, but also includes MRP and Greedy, and the mixed model condition. **(A)** Number of genes is fixed to 1000, and gene tree support is changed across different boxes by adjusting sequence length (from left to right, increasing sequence length); this produces datasets that have similar gene tree support as various subsets of the real avian dataset. **(B)** Gene tree support is fixed to the UCE-like case, and the number of genes is varied. **(C)** Gene tree support is fixed to the Intron-like case, and the the number of genes is varied. **(D)** A model condition that resembles the total evidence nucleotide dataset from the avian phylogenomics project with 14350 genes and mixed gene tree support: 8250 exons-like, 2500 intron-like, and 3600 UCE-like.

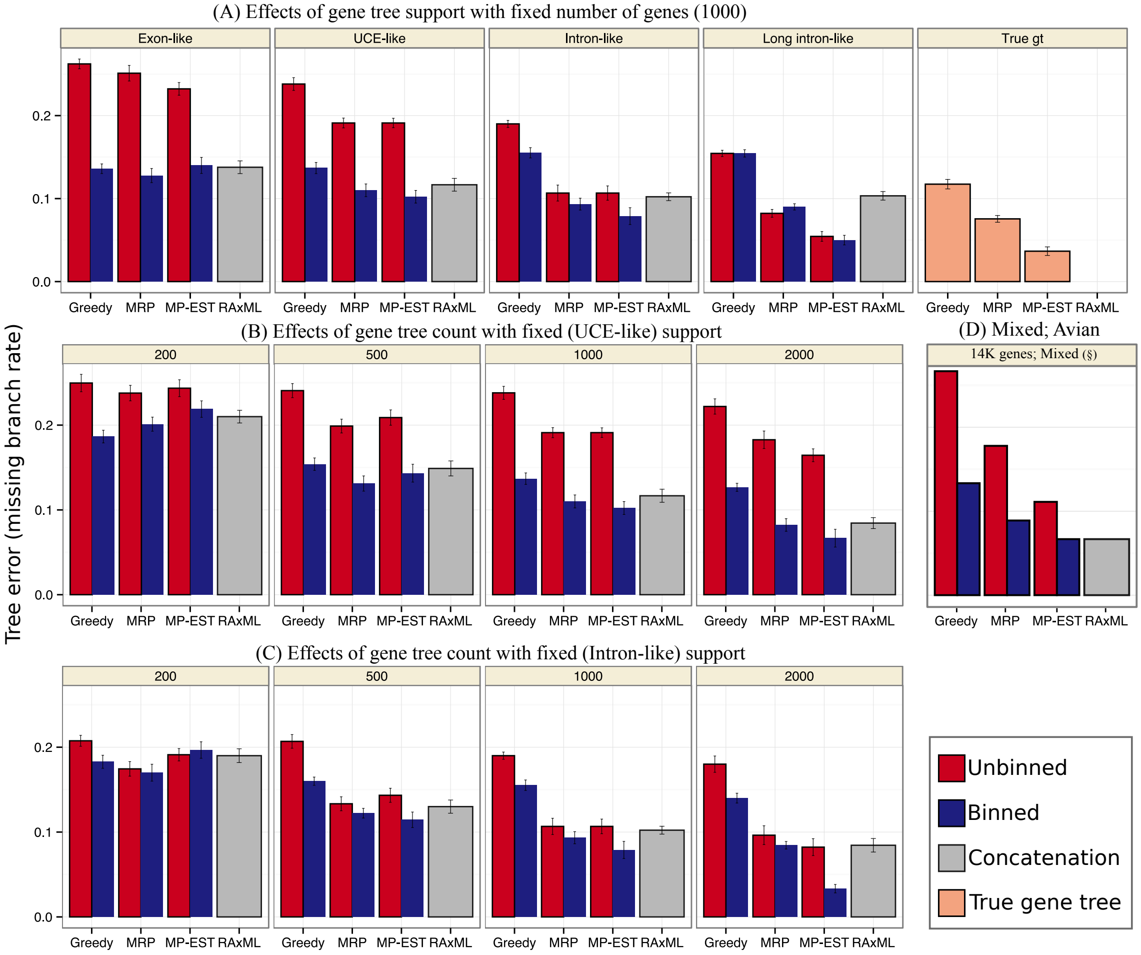


Figure S7: **False positive error rates on simulated avian datasets.** In some rare cases on the simulated avian datasets, the greedy consensus trees produced by the multi-locus bootstrapping procedure were missing one or two edges, and hence had small polytomies. In such cases, the missing branch (false negative) rate and false positive branch rates can be slightly different. For completeness, we show the false positive rates for all the cases shown in Figure 2 of the main paper and Figure S6. Bars show average false positive rates, over 20 replicates (except 2000 genes model conditions that had 10 replicates and mixed condition that had one replicate), and error-bars show standard error. Various panels show different model conditions: **(A)** 1000 genes with varying gene tree support, **(B)** varying number of gene trees with UCE-like support, **(C)** varying number of gene trees with Intron-like support, **(D)** 14350 genes with mixed support, resembling the total evidence avian dataset.

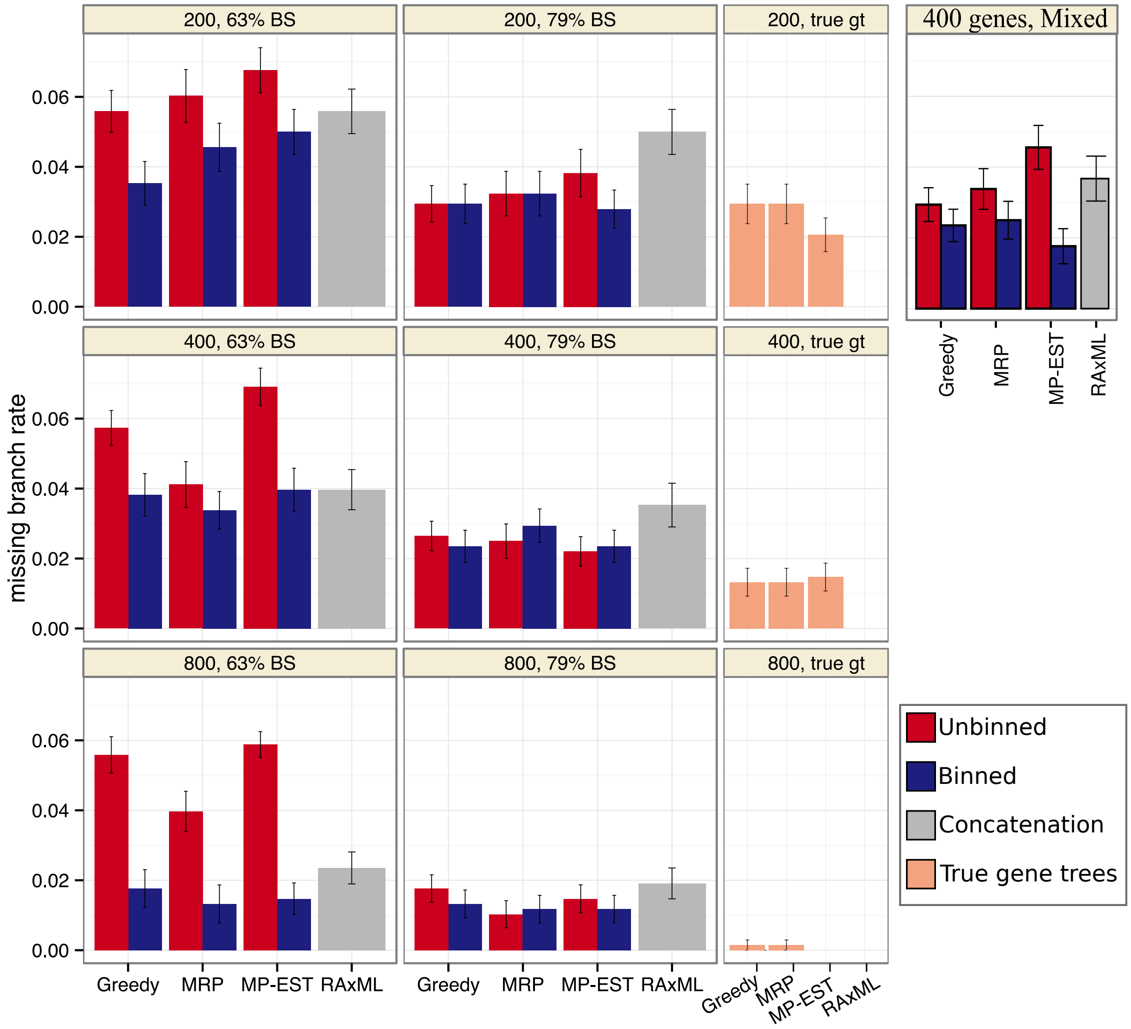


Figure S8: **Simulation results including MRP, Greedy, MP-EST, and concatenation using maximum likelihood on mammalian datasets.** Bars show average missing branch rate over 20 replicates, and error-bars show standard error. This figure shows similar results to Figure 3 of the main paper, but also includes MRP and Greedy. Different boxes correspond to various model conditions, with varying gene trees support (63% BS and 79% BS), and number of gene trees (200, 400, and 800 genes). We also show results on a mixed support condition with 200 63% BS gene trees and 200 79% BS gene trees. Thus, this model condition has 400 genes of average 71% BS, which closely resembles the biological mammalian dataset (424 gene trees of average 71% BS).

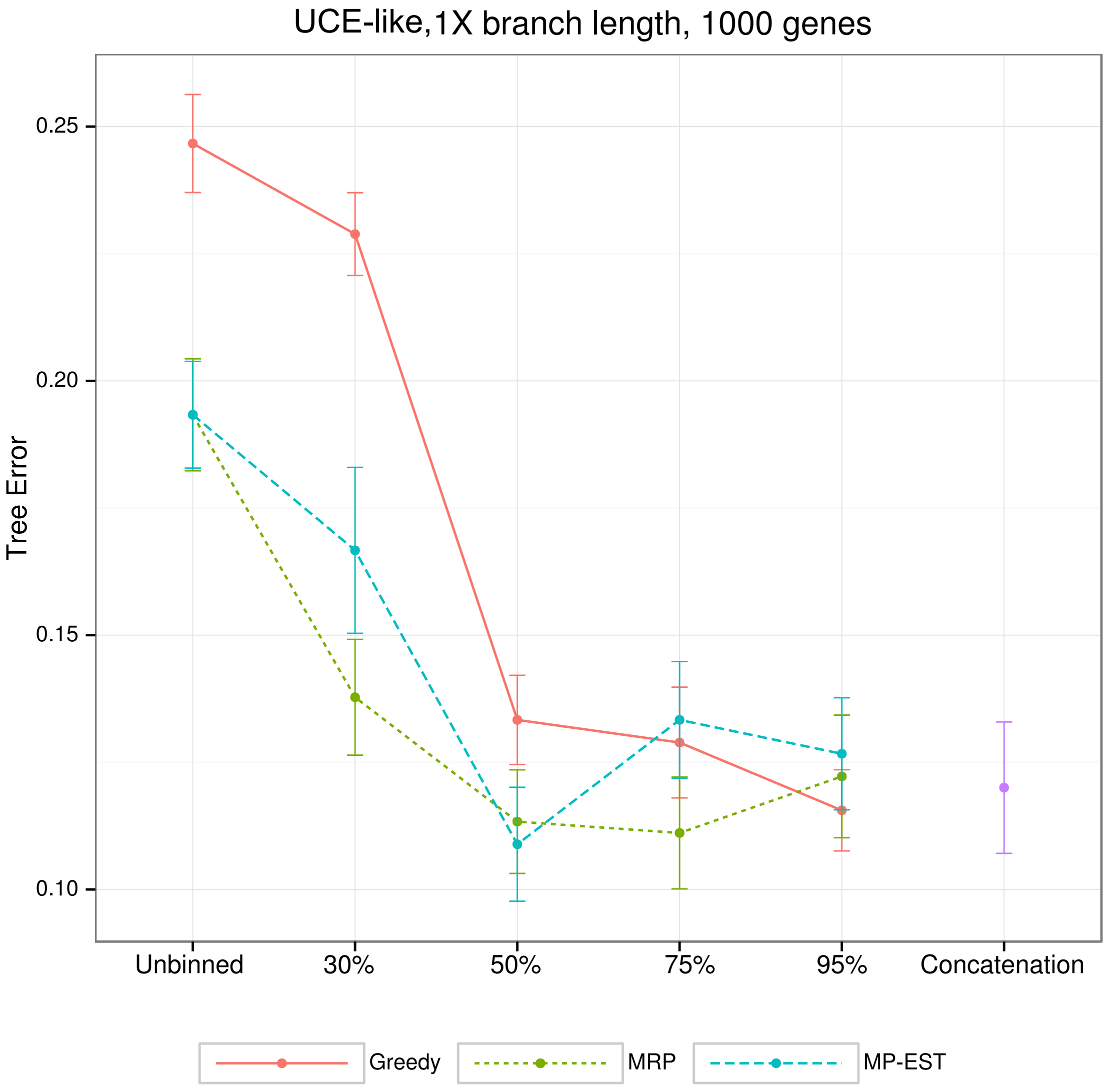


Figure S9: **Effects of support threshold t on the statistical binning.** Results are shown for the simulated avian with 10 replicates, 1000 genes, and UCE-like gene tree support. Dots correspond to average tree error and error bars correspond to standard error. Results are shown for unbinned analyses, binned analyses with 30%, 50%, 75%, and 95% support threshold, and concatenation. Both 50% and 75% thresholds give good results.

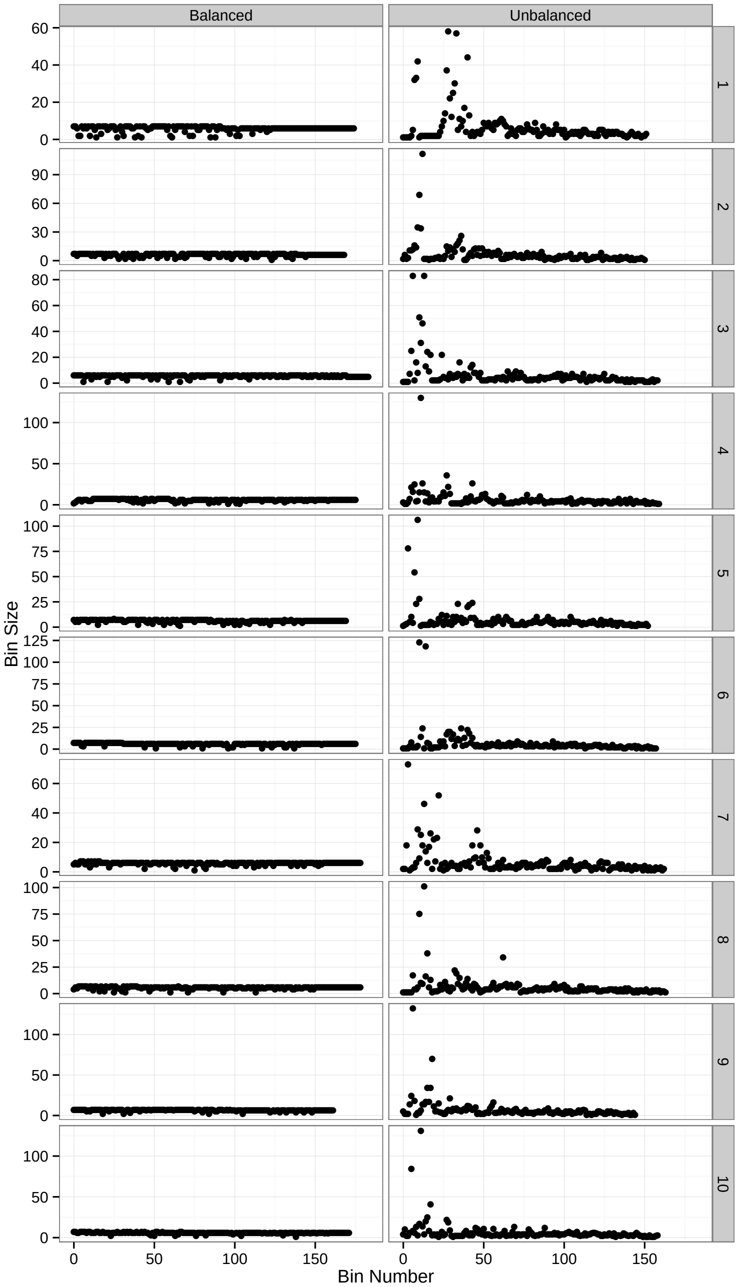


Figure S10: **Bin sizes on the simulated dataset produced by the original Brélaz heuristic ("unbalanced") and our modification ("balanced").** Results are shown for the first 10 replicates of the avian simulated UCE-like dataset with 1000 genes and $t = 50\%$. Each dot represents a bin, with vertical axis showing the bins size, and horizontal axis showing the bin index.

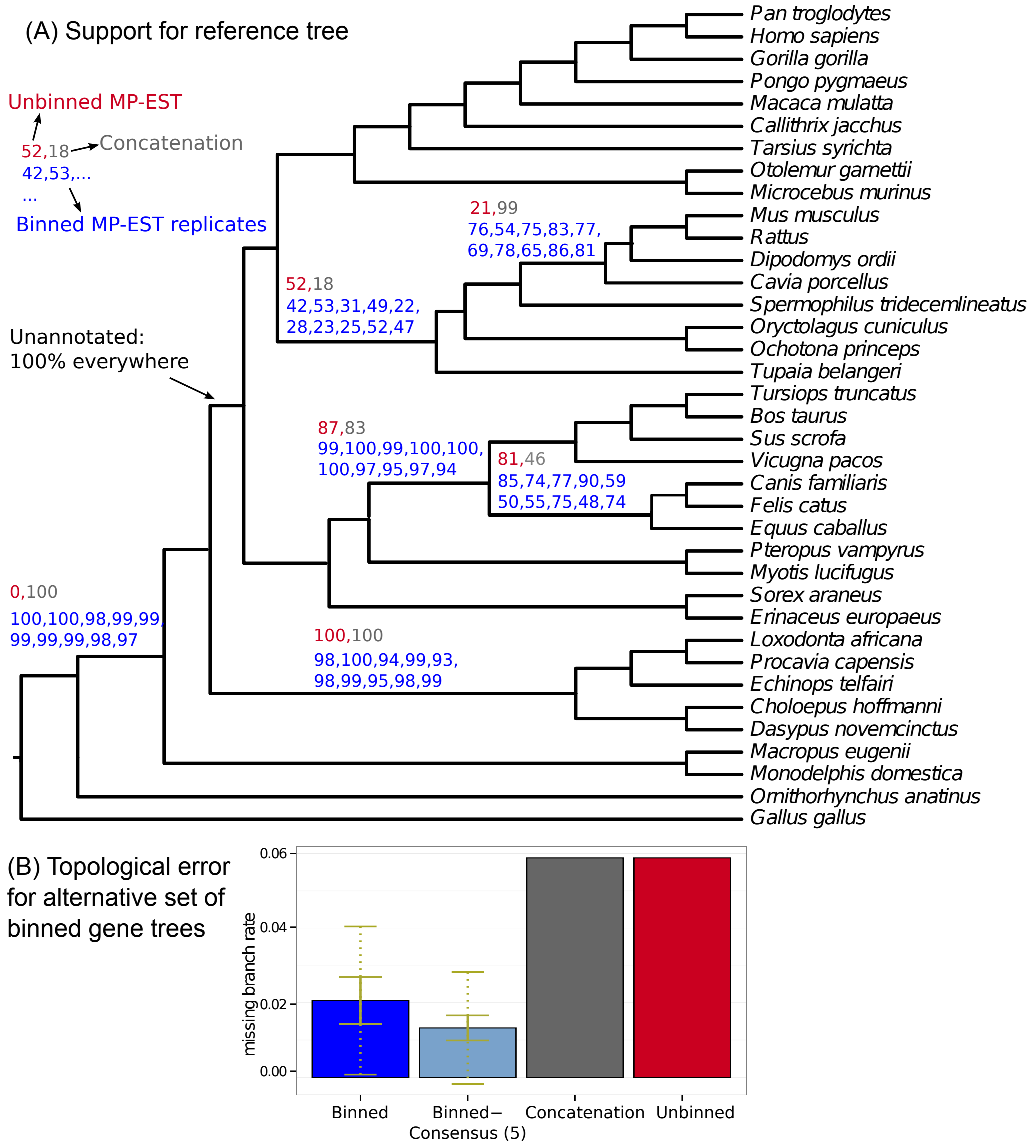


Figure S11: **Results for 10 different sets of bins obtained using alternative vertex colorings.** Our binning algorithm uses a greedy approach, in which it adds each gene to the *smallest* bin that does not conflict with it. In the presence of a tie (two or more smallest bins), one bin is randomly chosen. For the first replicate of the 63%BS mammalian dataset with 400 genes, we ran vertex coloring 10 times, breaking ties randomly. (A) Support for the reference tree using unbinned MP-EST (red), concatenation (grey), and each of the 10 binned MP-EST trees (blue). (B) Topological error for the 10 binned MP-EST trees, greedy consensus of 5 randomly selected (out of 10) binned MP-EST trees (20 replicates), unbinned MP-EST, and the concatenation tree. The error bars show the standard error (solid) and standard deviation (dotted). The binned MP-EST has better accuracy than both unbinned and concatenation, but there is some variation among the 10 replicates, and the greedy consensus of five runs has the best accuracy. Comparing multiple runs of the binning approach reveals potential impacts of breaking ties randomly.

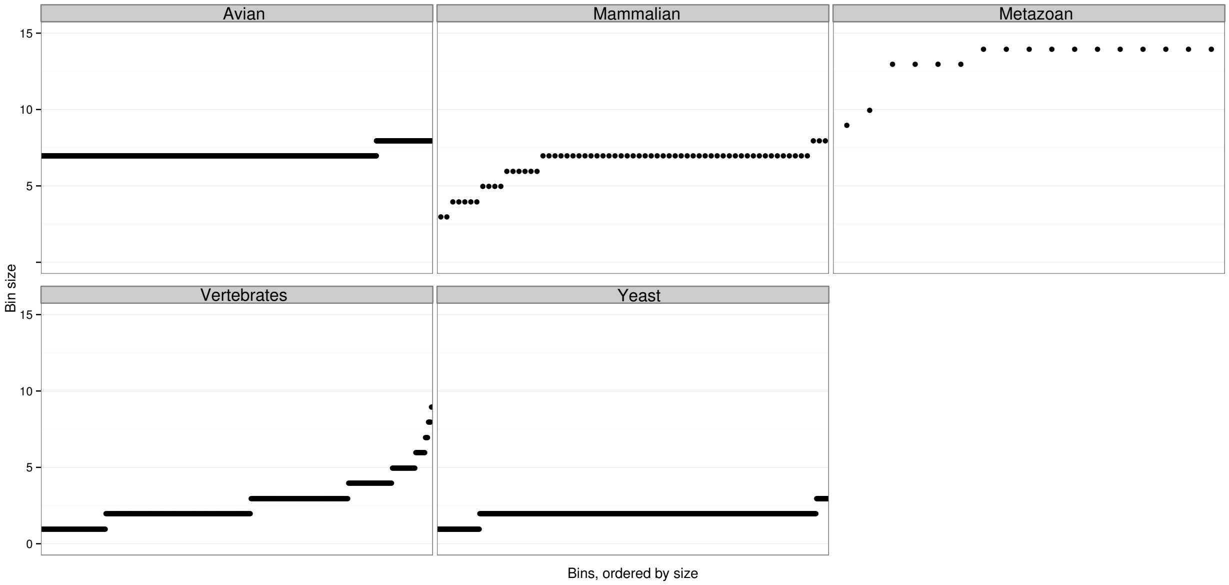


Figure S12: **Bin sizes for the five biological datasets.** On the avian and mammalian datasets almost all bins contain 7 genes (average is 7.1 for avian, and 6.5 for mammalian datasets); on the vertebrate dataset most bins are between 1 and 5 genes and the average is 2.7; and on the yeast dataset almost all bins have 2 genes. On the metazoan dataset, most bins are quite large, typically between 10 and 15 genes, and the average is 13.2 genes.

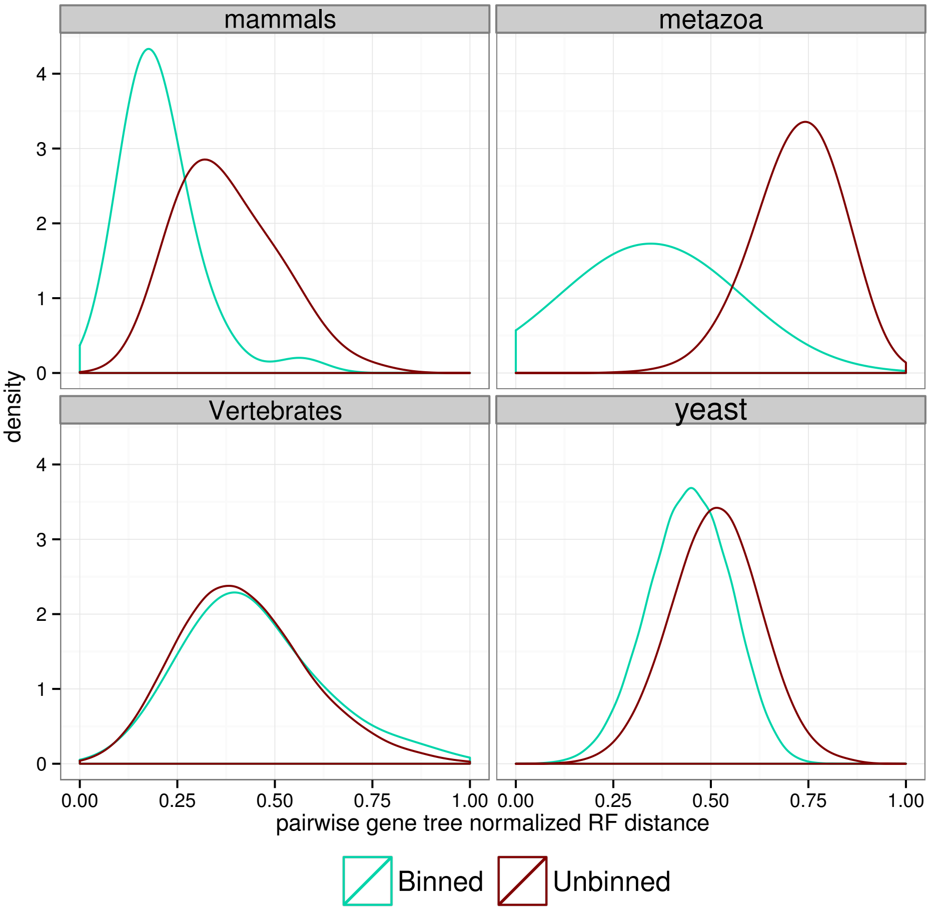


Figure S13: **Gene tree incongruence between pairs of estimated gene trees.** We measure gene tree incongruence using pairwise normalized RF distance between all pairs of estimated gene trees, with and without binning. For biological datasets, the distributions of pairwise gene tree distances are shown as kernel density plots. Except for the vertebrate dataset, where binning did not change the amount of incongruence between estimated gene trees, binning reduced incongruence between estimated gene trees.

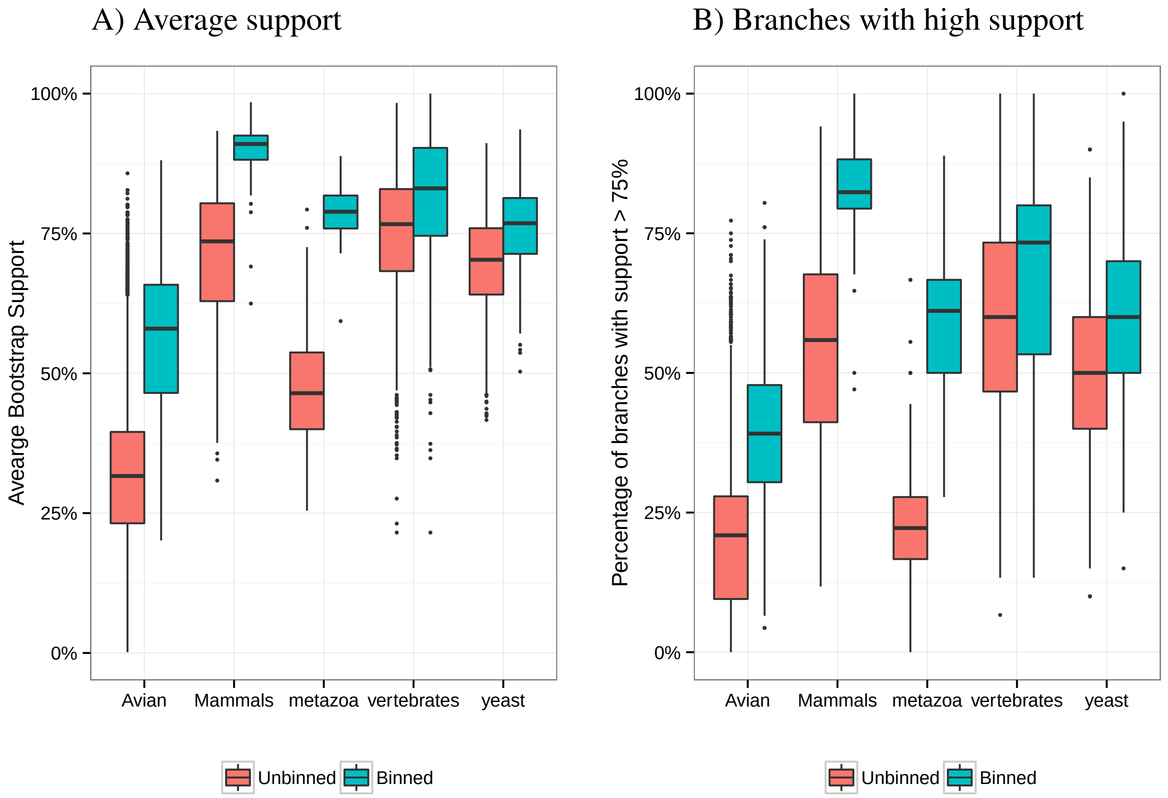


Figure S14: **Gene tree bootstrap support summaries for biological datasets.** A) Boxplots showing distribution of average bootstrap support across all estimated gene trees (solid lines indicate median values). B) Boxplots showing distribution of the percentage of branches in each gene tree that have support above 75% (solid lines indicate median values).

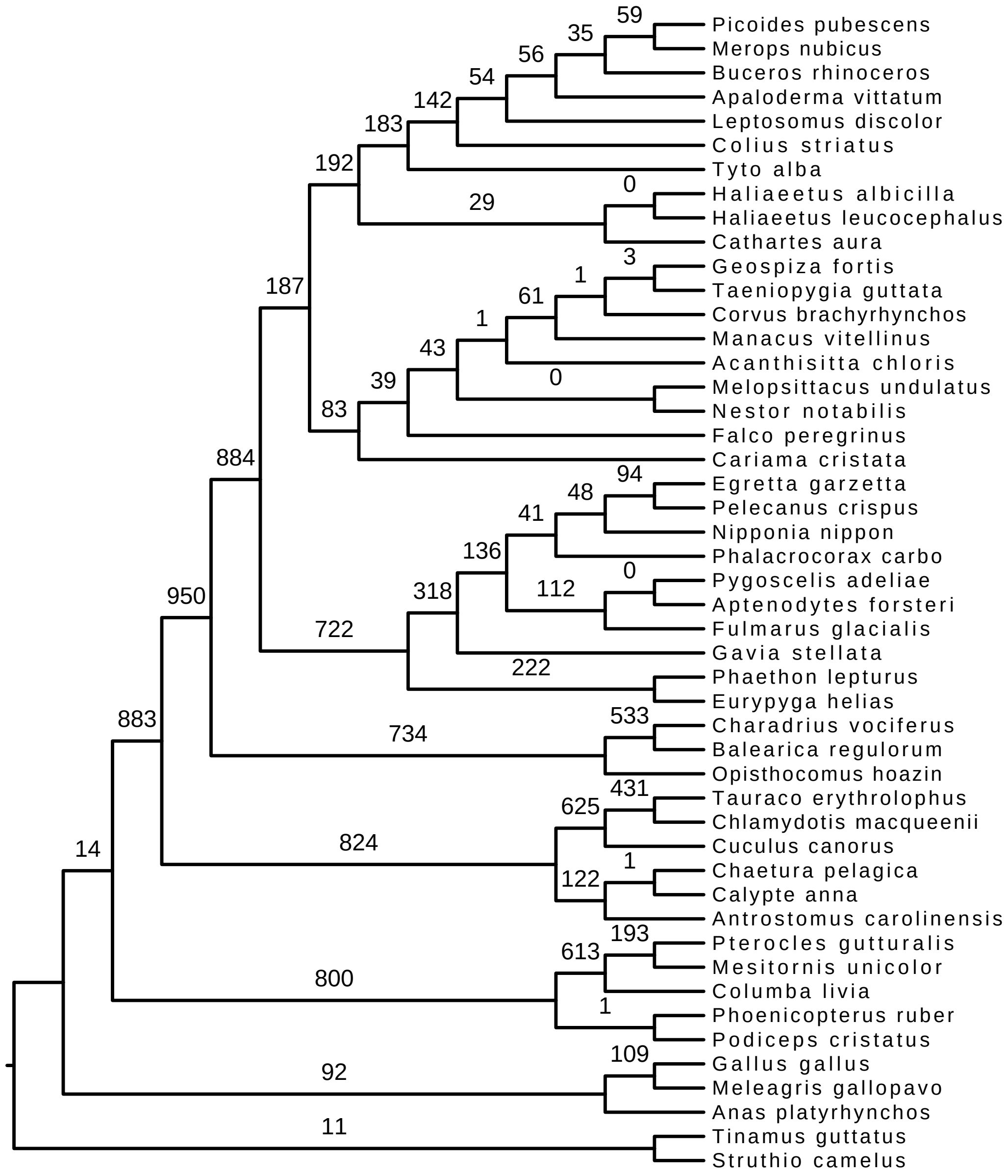


Figure S15: **Evidence for ILS in the avian dataset**. On each branch of the concatenation tree reported in (*31*), we show the number of intron gene trees (out of a total of 2516 loci) that rejected that branch with a BS of at least 75% (these results are also reported by (*31*)). Many of the edges of the tree are rejected by more than 200 gene trees with high support. Edges with lots of highly supported conflict are typically branches of the tree that are closer to the base, and also correspond to closely-spaced speciation events according to the dating analysis in (*31*), creating a very hard model condition.

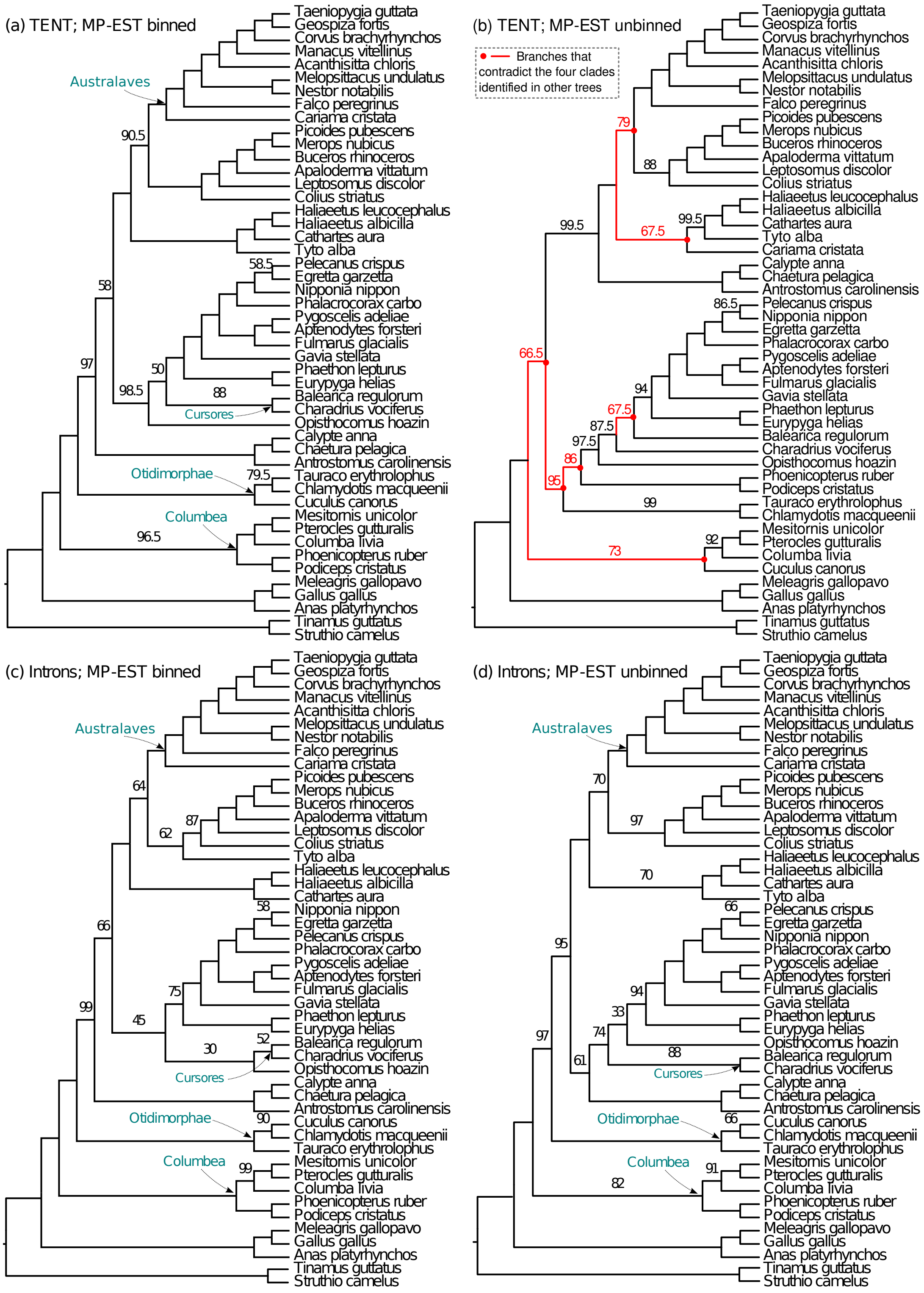


Figure S16: **Binned and unbinned MP-EST trees on the avian dataset.** All trees shown here are reported in (*31*).

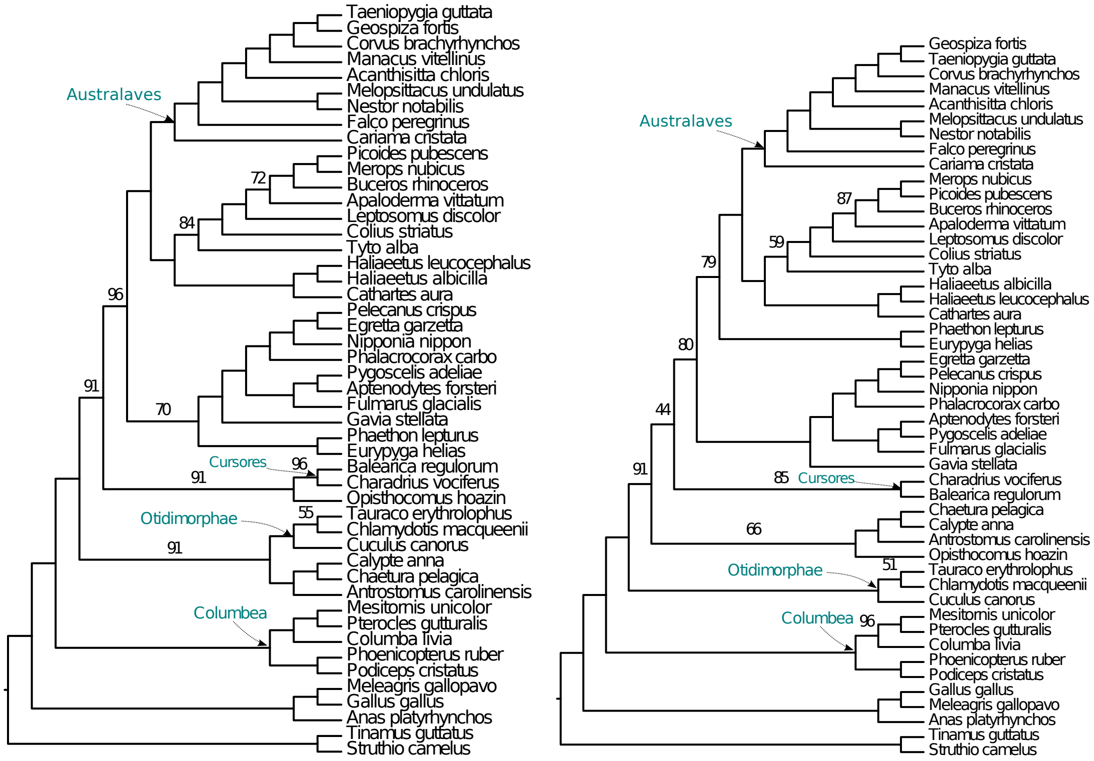


Figure S17: **The concatenation tree reported in (*31*) on the TENT (left) and intron (right) data matrices**. Branches without designation represent 100% support.

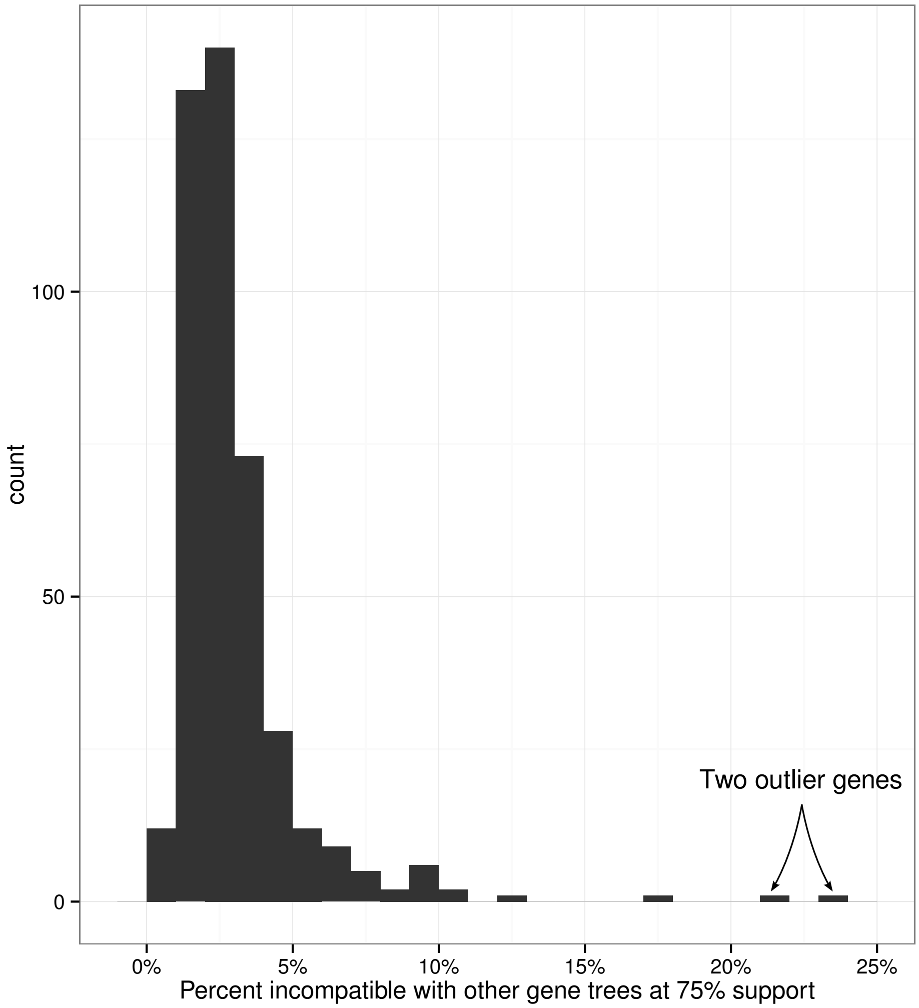
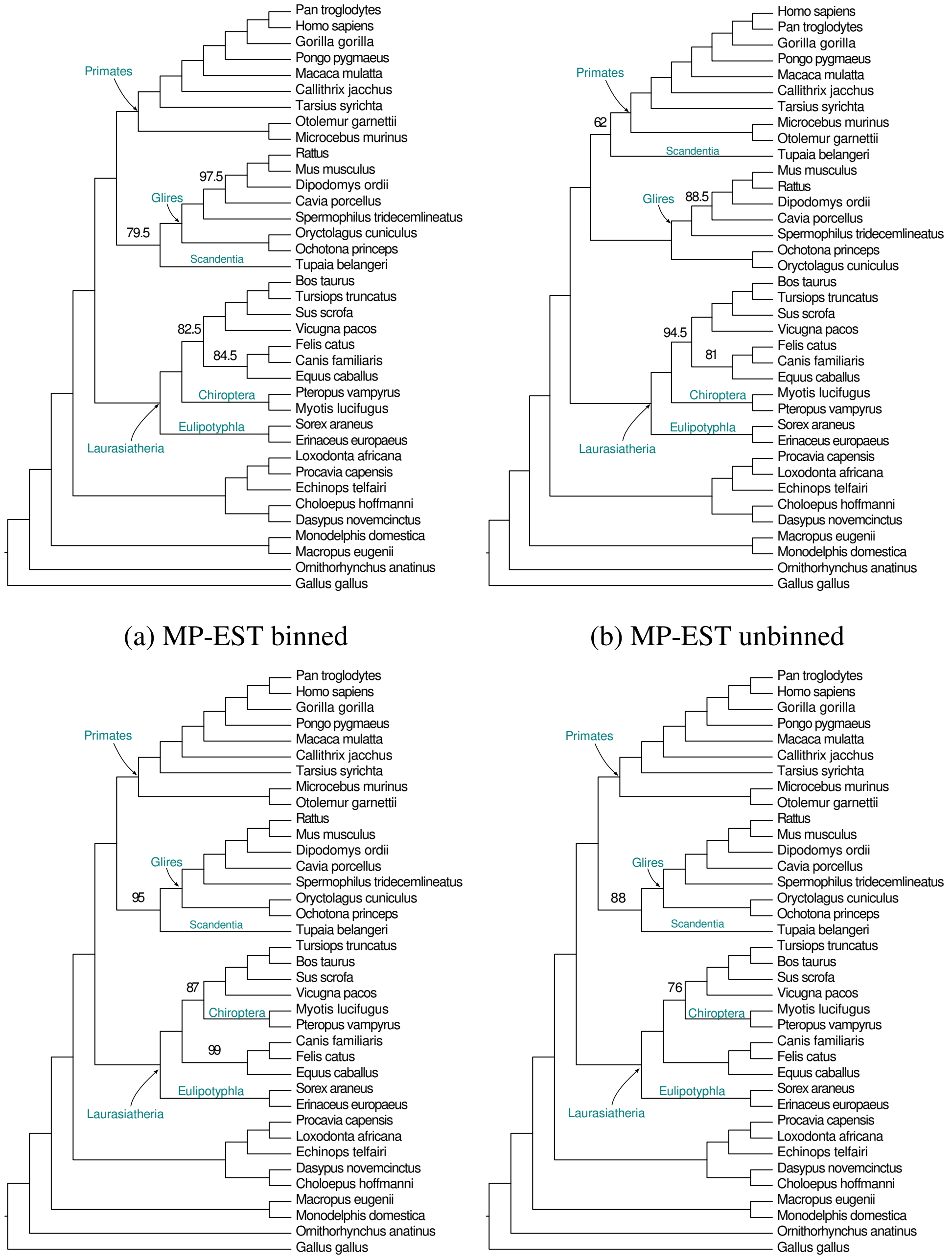


Figure S18: **Outlier mammalian genes.** The histogram shows the distribution of average percentage of branches that each gene had in conflict with other gene trees with at least 75% support. Two genes, with IDs 232 and 209, had on average more than 20% of their edges in conflict with other gene trees with bootstrap support higher than 75%. Including them in the binned analysis would have placed them in separate bins, and distorted the estimated triplet gene tree distribution. We also suspect these two gene trees have undetected estimation problems. Hence, we removed these two genes from the dataset.



(a) MP-EST binned

(b) MP-EST unbinned

(c) Concatenation; reported in (*6*)

(d) Concatenation; 424 genes

Figure S19: **Binned and unbinned MP-EST trees as well as concatenation trees on the reduced mammalian dataset.** We show results obtained on a reduced version of the mammalian dataset studied by Song *et al.* (*6*), after removing 23 problematic genes; we also show the original concatenated analysis by Song *et al.* on the full set of genes. Branches without designation have 100% support.

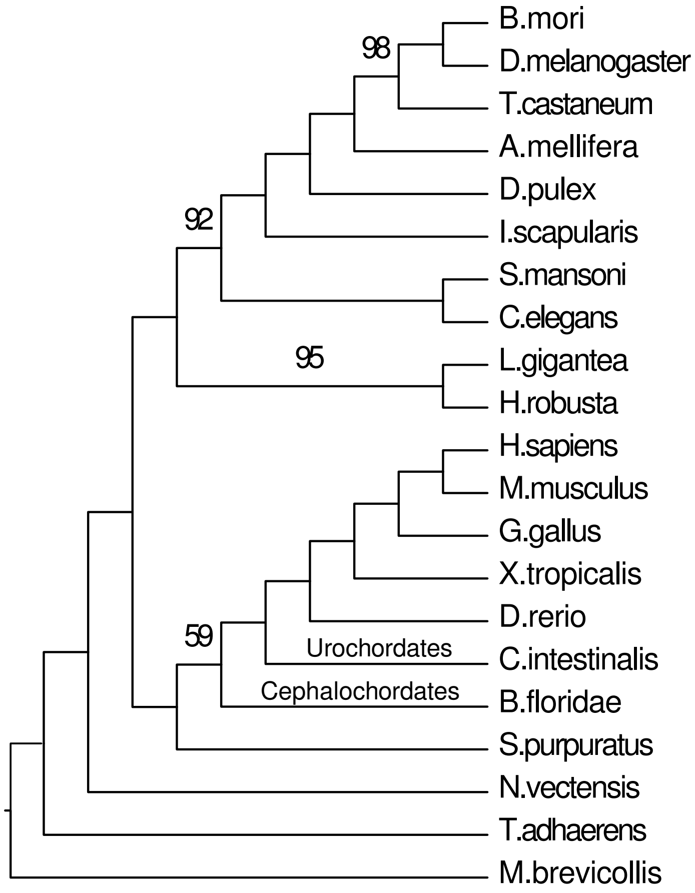


Figure S20: **The concatenation tree reported by Salichos and Rokas (*32*) on the metazoan dataset.** Branches without designation have 100% support. See the main text for MP-EST trees on the metazoa dataset.

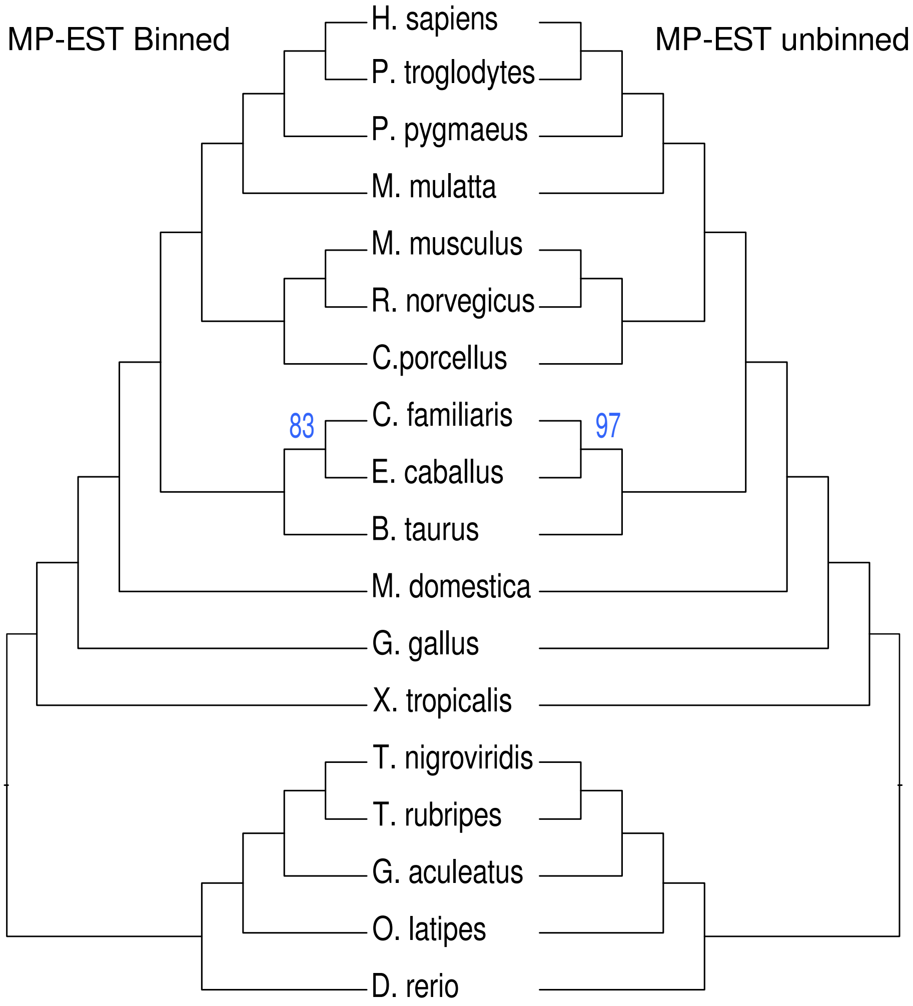
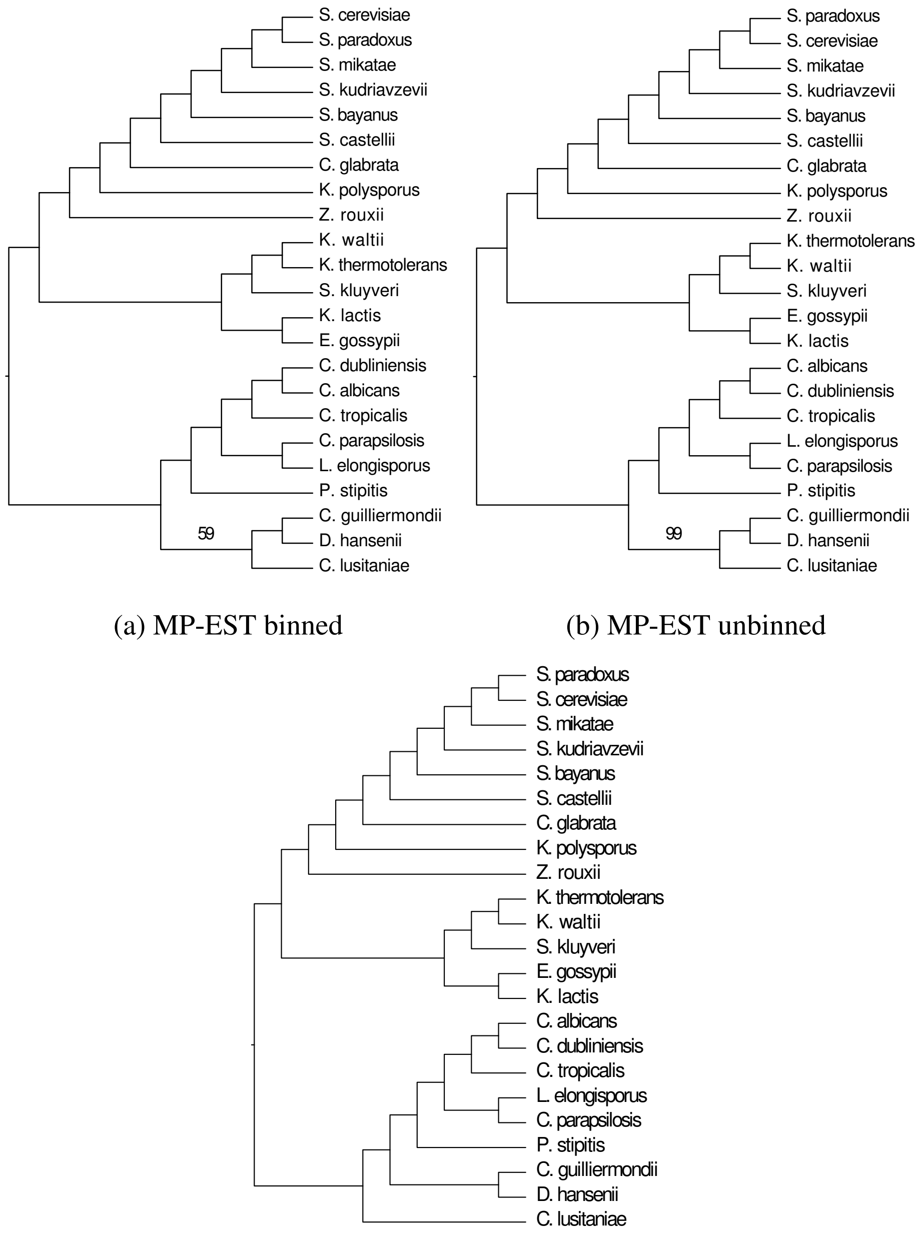


Figure S21: **Binned MP-EST and unbinned MP-EST on the vertebrates dataset.** Branches without designation have 100% support. The concatenation tree reported by Salichos and Rokas (*32*) is identical to these two trees and has 100% support for all branches.



(a) MP-EST binned

(b) MP-EST unbinned

(c) Concatenation; reported in (*32*)

Figure S22: **Results on the yeast biological dataset.** Binned and unbinned MP-EST trees are topologically identical, but differ in the bootstrap support for the branch uniting *C. lusitaniae* with *D. hansenii* and *C. guiliermondii*. This is the only branch with support below 100% and also the only edge that differs from the concatenation tree reported by Salichos and Rokas (*32*). The support on this branch in the unbinned analysis is very high (99%), but much lower (59%) in the binned analysis. Branches without designation have 100% support. The concatenation tree was reported to have 100% support for all edges.

4 Tables S1 to S6

	Mammals			**Avian**		
	2X	1X	0.5X	2X	1X	0.5X
Distance to Species Tree (mean)	18%	32%	54%	35%	47%	59%
Distance to Species Tree (min)	0%	3%	26%	16%	24%	36%
Distance to Species Tree (max)	42%	62%	82%	58%	69%	78%
Distance to other Gene Trees (mean)	26%	46%	71%	44%	57%	68%
Distance to other Gene Trees (min)	3%	14%	36%	16%	24%	38%
Distance to other Gene Trees (max)	51%	77%	96%	69%	77%	89%

Table S1: **True gene tree incongruence for simulated datasets, with varying model conditions.** 2X corresponds to the case where ILS is reduced by multiplying branch lengths by two and 0.5X corresponds to the case where ILS is increased by dividing branch lengths by two. The first three rows show average, minimum, and max normalized Robinson-Foulds (RF) distances between true gene trees and the model species tree. The next three rows show average, minimum, and maximum distances between all pairs of true gene trees. Maximum, minimum, and mean values shown are averages across all 20 replicates of 1000 genes for the avian dataset and 20 replicates of 200 genes for the mammals dataset.

(1000 genes, 1X ILS)		Exon-like	UCE-like	Intron-like	Long intron-like
Greedy	Unbinned	0.288 (0.024)	0.243 (0.030)	0.190 (0.027)	0.154 (0.017)
Greedy	Binned	0.139 (0.026)	0.138 (0.031)	0.158 (0.035)	0.154 (0.020)
MRP	Unbinned	0.251 (0.030)	0.191 (0.026)	0.107 (0.043)	0.082 (0.015)
MRP	Binned	0.128 (0.038)	0.110 (0.034)	0.093 (0.032)	0.090 (0.017))
MP-EST	Unbinned	0.232 (0.034)	0.191 (0.025)	0.107 (0.039)	0.054 (0.026)
MP-EST	Binned	0.140 (0.043)	**0.102 (0.034)**	**0.079 (0.045)**	**0.050 (0.026)**
RAxML	Concatenation	**0.138 (0.034)**	0.117 (0.034)	0.102 (0.021)	0.103 (0.023)
(UCE-like, 1X ILS)		200 genes	500 genes	1000 genes	2000 genes
Greedy	Unbinned	0.270 (0.040)	0.254 (0.032)	0.243 (0.030)	0.229 (0.030)
Greedy	Binned	**0.201 (0.031)**	0.163 (0.036)	0.138 (0.031)	0.127 (0.015)
MRP	Unbinned	0.238 (0.041)	0.199 (0.036)	0.191 (0.026)	0.183 (0.031)
MRP	Binned	**0.201 (0.037)**	**0.131 (0.040)**	0.110 (0.034)	0.082 (0.024)
MP-EST	Unbinned	0.244 (0.044)	0.209 (0.040)	0.191 (0.025)	0.164 (0.024)
MP-EST	Binned	0.219 (0.043)	0.143 (0.047)	**0.102 (0.034)**	**0.067 (0.033)**
RAxML	Concatenation	0.210 (0.033)	0.149 (0.040)	0.117 (0.034)	0.084 (0.020)
(Intron-like, 1X ILS)		200 genes	500 genes	1000 genes	2000 genes
Greedy	Unbinned	0.220 (0.027)	0.208 (0.035)	0.190 (0.027)	0.180 (0.030)
Greedy	Binned	0.196 (0.035)	0.166 (0.022)	0.158 (0.035)	0.140 (0.018)
MRP	Unbinned	0.174 (0.038)	0.133 (0.037)	0.107 (0.043)	0.096 (0.033)
MRP	Binned	**0.170 (0.045)**	0.122 (0.026)	0.093 (0.032)	0.084 (0.014)
MP-EST	Unbinned	0.191 (0.033)	0.143 (0.037)	0.107 (0.039)	0.082 (0.032)
MP-EST	Binned	0.197 (0.043)	**0.114 (0.041)**	**0.079 (0.045)**	**0.033 (0.016)**
RAxML	Concatenation	0.190 (0.036)	0.130 (0.035)	0.102 (0.021)	0.084 (0.025)
(True gene trees, 1X ILS)		200 genes	500 genes	1000 genes	2000 genes
Greedy	True gene tree	0.143 (0.026)	0.131 (0.024)	0.123 (0.025)	0.120 (0.021)
MRP	True gene tree	0.117 (0.034)	0.096 (0.024)	0.076 (0.018)	0.058 (0.012)
MP-EST	True gene tree	**0.110 (0.030)**	**0.053 (0.026)**	**0.037 (0.023)**	**0.018 (0.018)**
(1000 genes, UCE-like)		2X	1X	0.5X	
Greedy	Unbinned	0.226 (0.028)	0.243 (0.030)	0.293 (0.026)	
Greedy	Binned	0.077 (0.022)	0.138 (0.031)	0.262 (0.026)	
MRP	Unbinned	0.163 (0.019)	0.191 (0.026)	0.222 (0.042)	
MRP	Binned	**0.058 (0.017)**	0.110 (0.034)	0.174 (0.029)	
MP-EST	Unbinned	0.172 (0.030)	0.191 (0.025)	0.177 (0.044)	
MP-EST	Binned	0.059 (0.027)	**0.102 (0.034)**	**0.157 (0.045)**	
RAxML	Concatenation	0.064 (0.022)	0.117 (0.034)	0.197 (0.045)	
Greedy	True Gene Trees	0.066 (0.010)	0.123 (0.019)	0.181 (0.046)	
MRP	True Gene Trees	0.052 (0.011)	0.076 (0.019)	0.120 (0.024)	
MP-EST	True Gene Trees	**0.026 (0.018)**	**0.037 (0.019)**	**0.063 (0.030)**	

Table S2: **Missing branch rates for methods on the avian simulated dataset.** We show average missing branch rates and standard deviation in parentheses. Results are shown for various model conditions described in the paper, and the best method is shown in bold.

(63% BS; 1X ILS)		200 genes	400 genes	800 genes
Greedy	Unbinned	0.056 (0.027)	0.057 (0.022)	0.056 (0.023)
Greedy	Binned	**0.035 (0.028)**	0.038 (0.027)	0.018 (0.024)
MRP	Unbinned	0.060 (0.034)	0.041 (0.029)	0.040 (0.026)
MRP	Binned	0.046 (0.031)	**0.034 (0.024)**	**0.013 (0.024)**
MP-EST	Unbinned	0.068 (0.029)	0.069 (0.024)	0.059 (0.017)
MP-EST	Binned	0.050 (0.029)	0.040 (0.027)	0.015 (0.020)
RAxML	Concatenation	0.056 (0.028)	0.040 (0.026)	0.024 (0.020)
(79% BS; 1X ILS)		200 genes	400 genes	800 genes
Greedy	Unbinned	0.029 (0.023)	0.026 (0.019)	0.018 (0.018)
Greedy	Binned	0.029 (0.025)	0.024 (0.020)	0.013 (0.018)
MRP	Unbinned	0.032 (0.028)	0.025 (0.022)	**0.010 (0.017)**
MRP	Binned	0.032 (0.028)	0.029 (0.021)	0.012 (0.018)
MP-EST	Unbinned	0.038 (0.030)	**0.022 (0.019)**	0.015 (0.018)
MP-EST	Binned	**0.028 (0.024)**	0.024 (0.020)	0.012 (0.018)
RAxML	Concatenation	0.050 (0.029)	0.035 (0.028)	0.019 (0.020)
(True gene trees)		200 genes	400 genes	800 genes
Greedy	True gene trees	0.029 (0.025)	**0.013 (0.018)**	0.002 (0.007)
MRP	True gene trees	0.029 (0.025)	**0.013 (0.018)**	0.002 (0.007)
MP-EST	True gene trees	**0.021 (0.022)**	0.015 (0.018)	**0.000 (0.000)**
(63% BS; 200 genes)		2X	1X	0.5X
Greedy	Unbinned	0.049 (0.017)	0.056 (0.027)	0.074 (0.029)
Greedy	Binned	**0.019 (0.017)**	**0.035 (0.028)**	0.072 (0.024)
MRP	Unbinned	0.051 (0.021)	0.060 (0.034)	0.060 (0.028)
MRP	Binned	0.021 (0.019)	0.046 (0.031)	**0.059 (0.029)**
MP-EST	Unbinned	0.053 (0.018)	0.068 (0.029)	0.076 (0.028)
MP-EST	Binned	0.021 (0.022)	0.050 (0.029)	0.066 (0.030)
RAxML	Concatenation	0.024 (0.025)	0.056 (0.028)	0.078 (0.037)
Greedy	True gene trees	**0.003 (0.009)**	0.029 (0.025)	0.054 (0.027)
MRP	True gene trees	**0.003 (0.009)**	0.029 (0.025)	0.062 (0.036)
MP-EST	True gene trees	**0.003 (0.009)**	**0.021 (0.022)**	**0.051 (0.023)**

Table S3: **Missing branch rates of different methods on the mammalian simulated datasets.** We show average missing branch rates and standard deviations parenthetically. The top portion of the table shows results for the model conditions with 1X ILS level, gene tree support fixed to the 63% BS level and varying numbers of genes. Similarly, the middle portion shows results for 1X ILS level, fixing gene tree support to 79% BS level. The bottom portion shows results for the condition where number of genes is fixed to 200 genes, support to 63% BS, and the level of ILS is changed. All results are over 20 replicates and in for each model condition, the best method is shown in bold.

Fig.	Model condition	Binned vs. Unbinned	Binned vs. Concat.
	Avian simulated datasets		
2A	1X; 1000 genes, varying support	$\mathbf{p < 10^{-5}}$	$\mathbf{p = 0.00012}$
2B	1X; UCE-like, varying # genes	$\mathbf{p < 10^{-5}}$	$p = 0.37100$
2C	1X; Intron-like, varying # genes	$\mathbf{p = 0.00212}$	$\mathbf{p = 0.01058}$
4A	1k genes; UCE-like, varying ILS	$\mathbf{p < 10^{-5}}$	$\mathbf{p = 0.00418}$
	Mammalian simulated datasets		
3A	1X; 63% BS, varying # genes	$\mathbf{p < 10^{-5}}$	$p = 0.34300$
3A	1X: 79% BS, varying # genes	$p = 0.35690$	$\mathbf{p = 0.00315}$
4B	200 genes; 63% BS, varying ILS	$\mathbf{p = 0.00012}$	$p = 0.24818$

Table S4: **Statistical significance of binning on species tree topology for simulated datasets.** We evaluate the statistical significance of differences in species tree topology using an ANOVA test, with correction for multiple hypothesis using the Benjamini-Hochberg method (*70*) ($n = 14$), and setting $\alpha = 0.05$. For each model condition, two two-sided ANOVA tests are performed to establish whether binned MP-EST is better than unbinned MP-EST, and also whether binned MP-EST is better than concatenation. The two independent variables used in the ANOVA test are 1) the choice of the technique, and 2) the variable model parameter (e.g. support for Fig. 2A and number of genes for Fig. 2B). The p-values shown in the table are all for the first independent variable, i.e., the choice of the technique. The p-values for the interaction between the varying parameter and choice of the method are shown in Table S5. Differences between binned and unbinned MP-EST are statistically significant in all cases, except for 79% BS mammalian gene trees. Differences between concatenation and binned MP-EST are significant for four cases: Fig 2A (1000 avian genes of 1X ILS level, varying gene tree support), Fig 2C (Intronl-like avian genes of 1X ILS level, varying number of gene trees), Fig 4A (1000 avian genes of UCE-like support, varying ILS levels), and Fig 3A (mammalian genes with 79% BS and 1X ILS level, varying number of genes).

Fig.	Fixed Parameters	Variable	Binned vs. Unbinned	Binned vs. Concat.
		Avian simulated datasets		
2A	1X ILS; 1000 genes	support	$\mathbf{p < 10^{-5}}$	$\mathbf{p = 0.01162}$
2B	1X ILS; UCE-like	# genes	$\mathbf{p = 0.00330}$	$p = 0.56569$
2C	1X ILS; Intron-like	# genes	$p = 0.10614$	$p = 0.08705$
4A	1k genes; UCE-like	ILS	$\mathbf{p < 10^{-5}}$	$p = 0.15076$
		Mammalian simulated datasets		
3A	1X ILS; 63% BS	# genes	$p = 0.12071$	$p = 0.78500$
3A	1X ILS; 79% BS	# genes	$p = 0.56569$	$p = 0.50470$
4B	200 genes; 63% BS	ILS	$p = 0.25511$	$p = 0.78500$

Table S5: **Statistical significance of interaction between model parameters and performance of binning on species tree topology for simulated datasets.** Statistical significance of differences in species tree topology (dependent variable) are evaluated using a two-sided ANOVA test, with correction for multiple hypothesis using Benjamini Hochberg (*70*) ($n = 14$), and setting $\alpha = 0.05$. The two independent variables used in the ANOVA test are 1) the choice of the technique (Binned MP-EST vs. Unbinned MP-EST, and also Binned MP-EST vs. Concatenation), and 2) the variable model parameter (e.g. gene tree support for Fig. 2A and number of genes for Fig. 2B). Table S4 shows p-values for impact of the choice of the technique. Here, we show p-values for the interaction between the varying parameter and choice of the technique. Thus each p-value should be interpreted with regard to questions of the following form: "is the relative performance of binned MP-EST and unbinned MP-EST (or concatenation) affected by the choice of varying parameter." For example, for Fig. 2A, the p-value shown under Binned vs. Unbinned indicates that the gene tree support has a statistically significant impact on the relative performance of binned and unbinned MP-EST.

Avian Simulated Datasets (Figure 2)				
1000 genes; 1X (Fig. 2A)	Exon-like	UCE-like	Intron-like	Long intron-like
average bin size	15.1	5.8	1.8	1.2
UCE-like; 1X (Fig. 2B)	k=200	k=500	k=1000	k=2000
average bin size	4.2	5.0	5.8	6.6
Intron-like; 1X (Fig. 2C)	k=200	k=500	k=1000	k=2000
average bin size	1.6	1.7	1.8	2.0
Mixed; 1X (Fig. S6D)	k=14350			
average bin size	8.7			
Mammalian Simulated Datasets (Figure 3)				
63% BS; 1X (Fig. 3A)	k=200	k=400	k=800	
average bin size	2.5	2.8	3.2	
79% BS; 1X (Fig. 3A)	k=200	k=400	k=800	
average bin size	1.2	1.3	1.3	
Mixed; 1X (Fig. 3C)	k=400			
average bin size	1.8			
Impact of ILS (Figure 4)				
Avian; UCE-like (Fig. 4A)	2X	1X	0.5X	
average bin size	9.1	5.8	2.6	
Mammalian; 63% BS (Fig. 4B)	2X	1X	0.5X	
average bin size	4.9	2.5	1.2	

Table S6: **Average bin size for our statistical binning technique on different simulated datasets.** Results are shown for each dataset shown in Figure 2 (simulation results on the avian dataset), Figure 3 (simulation results on the mammalian dataset) and Figure 4 (simulation results showing effects of levels of ILS) of the main paper. The parameter k refers to the number of genes, "BS" refers to average bootstrap support value, and 1X, 2X, and 0.5X refer to the ILS levels. Thus 1X is observed levels of ILS, 2X is reduced ILS, and 0.5X is increased ILS.

References and Notes

1. A. Rokas, B. L. Williams, N. King, S. B. Carroll, Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798–804 (2003). Medline doi:10.1038/nature02053

2. W. P. Maddison, Gene trees in species trees. *Syst. Biol.* **46**, 523–536 (1997). doi:10.1093/sysbio/46.3.523

3. S. V. Edwards, Is a new and general theory of molecular systematics emerging? *Evolution* **63**, 1–19 (2009). Medline doi:10.1111/j.1558-5646.2008.00549.x

4. L. L. Knowles, Estimating species trees: Methods of phylogenetic analysis when there is incongruence across genes. *Syst. Biol.* **58**, 463–467 (2009). Medline doi:10.1093/sysbio/syp061

5. S. J. Hackett, R. T. Kimball, S. Reddy, R. C. Bowie, E. L. Braun, M. J. Braun, J. L. Chojnowski, W. A. Cox, K. L. Han, J. Harshman, C. J. Huddleston, B. D. Marks, K. J. Miglia, W. S. Moore, F. H. Sheldon, D. W. Steadman, C. C. Witt, T. Yuri, A phylogenomic study of birds reveals their evolutionary history. *Science* **320**, 1763–1768 (2008). Medline doi:10.1126/science.1157704

6. S. Song, L. Liu, S. V. Edwards, S. Wu, Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14942–14947 (2012). Medline doi:10.1073/pnas.1211733109

7. R. W. Meredith, J. E. Janečka, J. Gatesy, O. A. Ryder, C. A. Fisher, E. C. Teeling, A. Goodbla, E. Eizirik, T. L. Simão, T. Stadler, D. L. Rabosky, R. L. Honeycutt, J. J. Flynn, C. M. Ingram, C. Steiner, T. L. Williams, T. J. Robinson, A. Burk-Herrick, M. Westerman, N. A. Ayoub, M. S. Springer, W. J. Murphy, Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* **334**, 521–524 (2011). Medline doi:10.1126/science.1211028

8. J. H. Degnan, N. A. Rosenberg, Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24**, 332–340 (2009). Medline doi:10.1016/j.tree.2009.01.009

9. N. A. Rosenberg, Discordance of species trees with their most likely gene trees: A unifying principle. *Mol. Biol. Evol.* **30**, 2709–2713 (2013). Medline doi:10.1093/molbev/mst160

10. S. V. Edwards, L. Liu, D. K. Pearl, High-resolution species trees without concatenation. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 5936–5941 (2007). Medline doi:10.1073/pnas.0607004104

11. L. S. Kubatko, J. H. Degnan, Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* **56**, 17–24 (2007). Medline doi:10.1080/10635150601146041

12. B. R. Larget, S. K. Kotha, C. N. Dewey, C. Ané, BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* **26**, 2910–2911 (2010). Medline doi:10.1093/bioinformatics/btq539

13. L. Liu, L. Yu, S. V. Edwards, A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.* **10**, 302 (2010). Medline doi:10.1186/1471-2148-10-302

14. L. Liu, D. K. Pearl, Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* **56**, 504–514 (2007). Medline doi:10.1080/10635150701429982

15. L. Liu, BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* **24**, 2542–2543 (2008). Medline doi:10.1093/bioinformatics/btn484

16. L. Liu, L. Yu, D. K. Pearl, S. V. Edwards, Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* **58**, 468–477 (2009). Medline doi:10.1093/sysbio/syp031

17. E. Mossel, S. Roch, Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci. *IEEE/ACM Trans. Comp. Biol. Bioinform.* **7**, 166–171 (2010).doi:10.1109/TCBB.2008.66 Medline

18. J. Heled, A. J. Drummond, Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* **27**, 570–580 (2010). Medline doi:10.1093/molbev/msp274

19. Y. Yu, T. Warnow, L. Nakhleh, Algorithms for MDC-based multi-locus phylogeny inference: Beyond rooted binary gene trees on single alleles. *J. Comput. Biol.* **18**, 1543–1559 (2011). Medline doi:10.1089/cmb.2011.0174

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

21. D. Bryant, R. Bouckaert, J. Felsenstein, N. A. Rosenberg, A. RoyChoudhury, Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* **29**, 1917–1932 (2012). Medline doi:10.1093/molbev/mss086

22. N. De Maio, C. Schlötterer, C. Kosiol, Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Mol. Biol. Evol.* **30**, 2249–2262 (2013). Medline doi:10.1093/molbev/mst131

23. J. Chifman, L. Kubatko, Quartet inference from SNP data under the coalescent model. *Bioinformatics* 10.1093/bioinformatics/btu530 (2014).doi: 10.1093/bioinformatics/btu530

24. B. T. Smith, M. G. Harvey, B. C. Faircloth, T. C. Glenn, R. T. Brumfield, Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Syst. Biol.* **63**, 83–95 (2014). Medline doi:10.1093/sysbio/syt061

25. R. T. Kimball, N. Wang, V. Heimer-McGinn, C. Ferguson, E. L. Braun, Identifying localized biases in large datasets: A case study using the avian tree of life. *Mol. Phylogenet. Evol.* **69**, 1021–1032 (2013). Medline doi:10.1016/j.ympev.2013.05.029

26. J. E. McCormack, M. G. Harvey, B. C. Faircloth, N. G. Crawford, T. C. Glenn, R. T. Brumfield, A phylogeny of birds based on over 1,500 loci collected by target enrichment

and high-throughput sequencing. *PLOS ONE* **8**, e54848 (2013). Medline doi:10.1371/journal.pone.0054848

27. M. DeGiorgio, J. H. Degnan, Fast and consistent estimation of species trees using supermatrix rooted triples. *Mol. Biol. Evol.* **27**, 552–569 (2010). Medline doi:10.1093/molbev/msp250

28. M. S. Bayzid, T. Warnow, Naive binning improves phylogenomic analyses. *Bioinformatics* **29**, 2277–2284 (2013). Medline doi:10.1093/bioinformatics/btt394

29. S. Patel, R. Kimball, E. Braun, Error in phylogenetic estimation for bushes in the Tree of Life. *J. Phylogenet. Evol. Biol.* **1**, 2 (2013). doi:10.4172/2329-9002.1000110

30. L. Nakhleh, U. Roshan, K. St. John, J. Sun, T. Warnow, Designing fast converging phylogenetic methods. *Bioinformatics* **17** (suppl. 1), S190–S198 (2001). Medline doi:10.1093/bioinformatics/17.suppl_1.S190

31. E. Jarvis, S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, A. Suh, C. C. Weber, R. R. da Fonseca, J. Li, F. Zhang, H. Li, L. Zhou, N. Narula, L. Liu, G. Ganapathy, B. Boussau, Md. S. Bayzid, V. Zavidovych, S. Subramanian, T. Gabaldón, S. Capella-Gutiérrez, J. Huerta-Cepas, B. Rekepalli, K. Munch, M. Schierup, B. Lindow, W. C. Warren, D. Ray, R. E. Green, M. Bruford, X. Zhan, A. Dixon, S. Li, N. Li, Yinhua Huang, E. P. Derryberry, M. Frost Bertelsen, F. H. Sheldon, R. T. Brumfield, C. V. Mello, P. V. Lovell, M. Wirthlin, M. P. C. Schneider, F. Prosdocimi, J. A. Samaniego, A. M. Vargas Velazquez, A. Alfaro-Núñez, P. F. Campos, B. Petersen, T. Sicheritz-Ponten, A. Pas, T. Bailey, P. Scofield, M. Bunce, D. M. Lambert, Q. Zhou, P. Perelman, A. C. Driskell, B. Shapiro, Z. Xiong, Y. Zeng, S. Liu, Z. Li, B. Liu, K. Wu, J. Xiao, X. Yinqi, Q. Zheng, Y. Zhang, H. Yang, J. Wang, L. Smeds, F. E. Rheindt, M. Braun, J. Fjeldsa, L. Orlando, K. Barker, K. A. Jønsson, W. Johnson, K.-P. Koepfli, S. O'Brien, D. Haussler, O. A. Ryder, C. Rahbek, E. Willerslev, G. R. Graves, T. C. Glenn, J. McCormack, D. Burt, H. Ellegren, P. Alström, S. V. Edwards, A. Stamatakis, D. P. Mindell, J. Cracraft, E. L. Braun, T. Warnow, W. Jun, M. T. P. Gilbert, G. Zhang, Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 10.1126/science.1253451 [this issue] (2013).

32. L. Salichos, A. Rokas, Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327–331 (2013). Medline doi:10.1038/nature12130

33. A. Stamatakis, RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006). Medline doi:10.1093/bioinformatics/btl446

34. Materials and methods are available as supplementary material on *Science* Online.

35. M. A. Ragan, Phylogenetic inference based on matrix representation of trees. *Mol. Phylogenet. Evol.* **1**, 53–58 (1992). Medline doi:10.1016/1055-7903(92)90035-F

36. T.-K. Seo, Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Mol. Biol. Evol.* **25**, 960–971 (2008). Medline doi:10.1093/molbev/msn043

37. A. D. Leaché, R. B. Harris, B. Rannala, Z. Yang, The influence of gene flow on species tree estimation: A simulation study. *Syst. Biol.* **63**, 17–30 (2013). Medline doi:10.1093/sysbio/syt049

38. L. Zhao, N. Zhang, P. F. Ma, Q. Liu, D. Z. Li, Z. H. Guo, Phylogenomic analyses of nuclear genes reveal the evolutionary relationships within the BEP clade and the evidence of positive selection in Poaceae. *PLOS ONE* **8**, e64642 (2013). Medline doi:10.1371/journal.pone.0064642

39. B. Zhong, L. Liu, Z. Yan, D. Penny, Origin of land plants using the multispecies coalescent model. *Trends Plant Sci.* **18**, 492–495 (2013). Medline doi:10.1016/j.tplants.2013.04.009

40. J. H. Degnan, M. DeGiorgio, D. Bryant, N. A. Rosenberg, Properties of consensus methods for inferring species trees from gene trees. *Syst. Biol.* **58**, 35–54 (2009). Medline doi:10.1093/sysbio/syp008

41. A. Suh, M. Paus, M. Kiefmann, G. Churakov, F. A. Franke, J. Brosius, J. O. Kriegs, J. Schmitz, Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. *Nat. Commun.* **2**, 443 (2011). Medline doi:10.1038/ncomms1448

42. N. Wang, E. L. Braun, R. T. Kimball, Testing hypotheses about the sister group of the passeriformes using an independent 30-locus data set. *Mol. Biol. Evol.* **29**, 737–750 (2012). Medline doi:10.1093/molbev/msr230

43. C. Nielsen, *Animal Evolution: Interrelationships of the Living Phyla* (Oxford Univer Press, Oxford, 2012).

44. F. Delsuc, H. Brinkmann, D. Chourrout, H. Philippe, Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965–968 (2006). Medline doi:10.1038/nature04336

45. S. J. Bourlat, T. Juliusdottir, C. J. Lowe, R. Freeman, J. Aronowicz, M. Kirschner, E. S. Lander, M. Thorndyke, H. Nakano, A. B. Kohn, A. Heyland, L. L. Moroz, R. R. Copley, M. J. Telford, Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* **444**, 85–88 (2006). Medline doi:10.1038/nature05241

46. T. R. Singh, G. Tsagkogeorga, F. Delsuc, S. Blanquart, N. Shenkar, Y. Loya, E. J. Douzery, D. Huchon, Tunicate mitogenomics and phylogenetics: Peculiarities of the *Herdmania momus* mitochondrial genome and support for the new chordate phylogeny. *BMC Genomics* **10**, 534 (2009). Medline doi:10.1186/1471-2164-10-534

47. J. E. Janecka, W. Miller, T. H. Pringle, F. Wiens, A. Zitzmann, K. M. Helgen, M. S. Springer, W. J. Murphy, Molecular and genomic data identify the closest living relative of primates. *Science* **318**, 792–794 (2007). Medline doi:10.1126/science.1147555

48. B. Boussau, G. J. Szöllosi, L. Duret, M. Gouy, E. Tannier, V. Daubin, Genome-scale coestimation of species and gene trees. *Genome Res.* **23**, 323–330 (2013). Medline doi:10.1101/gr.141978.112

49. H. C. Lanier, L. L. Knowles, Is recombination a problem for species-tree analyses? *Syst. Biol.* **61**, 691–701 (2012). Medline doi:10.1093/sysbio/syr128

50. K. Liu, S. Raghavan, S. Nelesen, C. R. Linder, T. Warnow, Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* **324**, 1561–1564 (2009). Medline doi:10.1126/science.1171243

51. J. Felsenstein, Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401 (1978). doi:10.2307/2412923

52. W. G. Weisburg, S. J. Giovannoni, C. R. Woese, The Deinococcus-Thermus phylum and the effect of rRNA composition on phylogenetic tree construction. *Syst. Appl. Microbiol.* **11**, 128–134 (1989). Medline doi:10.1016/S0723-2020(89)80051-7

53. M. DeGiorgio, J. H. Degnan, Robustness to divergence time underestimation when inferring species trees from estimated gene trees. *Syst. Biol.* **63**, 66–82 (2014). Medline doi:10.1093/sysbio/syt059

54. J. W. Leigh, E. Susko, M. Baumgartner, A. J. Roger, Testing congruence in phylogenomic analysis. *Syst. Biol.* **57**, 104–115 (2008).doi:10.1080/10635150801910436 Medline

55. T. Warnow, Tree compatibility and inferring evolutionary history. *J. Algorithms* **16**, 388–407 (1994). doi:10.1006/jagm.1994.1018

56. D. Brélaz, New methods to color the vertices of a graph. *Commun. ACM* **22**, 251–256 (1979). doi:10.1145/359094.359101

57. T. P. Wilcox, D. J. Zwickl, T. A. Heath, D. M. Hillis, Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. *Mol. Phylogenet. Evol.* **25**, 361–371 (2002). Medline doi:10.1016/S1055-7903(02)00244-0

58. J. Sukumaran, M. T. Holder, DendroPy: A Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571 (2010). Medline doi:10.1093/bioinformatics/btq228

59. J. Dutheil, B. Boussau, Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol. Biol.* **8**, 255 (2008). Medline doi:10.1186/1471-2148-8-255

60. S. Kullback, R. A. Leibler, On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951). doi:10.1214/aoms/1177729694

61. D. Swofford, *PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods)* Version 4 (Sinauer Associates, Sunderland, MA, (2002).

62. B. Schierwater, M. Eitel, W. Jakob, H.-J. Osigus, H. Hadrys, S. L. Dellaporta, S.-O. Kolokotronis, R. Desalle, Concatenated analysis sheds light on early metazoan evolution and fuels a modern “urmetazoon” hypothesis. *PLOS Biol.* **7**, e20 (2009).doi:10.1371/journal.pbio.1000020 Medline

63. H. Philippe, R. Derelle, P. Lopez, K. Pick, C. Borchiellini, N. Boury-Esnault, J. Vacelet, E. Renard, E. Houliston, E. Quéinnec, C. Da Silva, P. Wincker, H. Le Guyader, S. Leys, D. J. Jackson, F. Schreiber, D. Erpenbeck, B. Morgenstern, G. Wörheide, M. Manuel, Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.* **19**, 706–712 (2009). Medline doi:10.1016/j.cub.2009.02.052

64. J. F. Ryan, K. Pang, C. E. Schnitzler, A.-D. Nguyen, R. T. Moreland, D. K. Simmons, B. J. Koch, W. R. Francis, P. Havlak, S. A. Smith, N. H. Putnam, S. H. D. Haddock, C. W. Dunn, T. G. Wolfsberg, J. C. Mullikin, M. Q. Martindale, A. D. Baxevanis; NISC Comparative Sequencing Program, The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science* **342**, 1242592 (2013). doi:10.1126/science.1242592 Medline

65. G. D. Edgecombe, G. Giribet, C. W. Dunn, A. Hejnol, R. M. Kristensen, R. C. Neves, G. W. Rouse, K. Worsaae, M. V. Sørensen, Higher-level metazoan relationships: Recent

progress and remaining questions. *Org. Divers. Evol.* **11**, 151–172 (2011). doi:10.1007/s13127-011-0044-4

66. K. Ishiwata, G. Sasaki, J. Ogawa, T. Miyata, Z.-H. Su, Phylogenetic relationships among insect orders based on three nuclear protein-coding gene sequences. *Mol. Phylogenet. Evol.* **58**, 169–180 (2011). Medline doi:10.1016/j.ympev.2010.11.001

67. J.-Y. Hu, Y.-P. Zhang, L. Yu, Summary of Laurasiatheria (mammalia) phylogeny. *Dongwuxue Yanjiu* **33** (E5-6), E65–E74 (2012). Medline doi:10.3724/SP.J.1141.2012.E05-06E65

68. B. Dujon, Yeast evolutionary genomics. *Nat. Rev. Genet.* **11**, 512–524 (2010). Medline doi:10.1038/nrg2811

69. M. Hubert, E. Vandervieren, An adjusted boxplot for skewed distributions. *Comput. Stat. Data Anal.* **52**, 5186–5201 (2008). doi:10.1016/j.csda.2007.11.008

70. Y. Benjamini, Y. Hochberg, *J. R. Statist. Soc. B* **57**, 289 (1995).