Statistical binning improves species tree estimation in the presence of gene tree incongruence

Siavash Mirarab,[1] Md Shamsuzzoha Bayzid,[1] Bastien Boussau,[2] Tandy Warnow[1*]

[1]Department of Computer Science, University of Texas at Austin,
[2]Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, France

*To whom correspondence should be addressed; E-mail: tandy@cs.utexas.edu

Gene tree incongruence resulting from incomplete lineage sorting (ILS) can make standard species tree estimation methods, such as concatenation, have poor accuracy. While coalescent-based methods can accurately estimate species trees in the presence of ILS, these methods can be less accurate than concatenation when estimated gene trees are insufficiently accurate. We present a new "statistical binning" technique to address this challenge: we use a statistical test for combinability and a graph-theoretic optimization to bin genes into subsets, estimate trees on each subset using concatenation, and then combine the "supergene" trees using the preferred coalescent-based method. We show that statistical binning dramatically improves the accuracy of MP-EST, a leading coalescent-based method, and we use statistical binning to produce the first genome-scale coalescent-based avian tree of life.

Introduction

Species trees provide a basis for understanding how life evolved on earth, as well as having applications to comparative genomics, orthology detection, protein function inference, and biodiversity analysis. Estimations of species trees are typically based on multiple genes, in some cases from throughout the genome. One advantage of such a "phylogenomic" approach is that it enables more data to be used to inform the tree estimation (*1*). However, there is increasing evidence that genes can have conflicting evolutionary histories due to many biological causes, including duplication and loss, horizontal gene transfer, recombination, and incomplete lineage sorting (ILS), and further complicated by incorrect orthology assessments. ILS is especially common in the presence of short branches or large population sizes, and high levels of ILS are expected in rapid radiations (such as is believed to have occurred in the avian and mammalian histories) (*2–6*). However, the standard phylogenetic estimation technique of concatenation, which operates by concatenating alignments for individual genes into a "combined" dataset and then estimates the species tree from the super-alignment, can sometimes return incorrect species trees with high confidence when there is substantial ILS (*7–10*).

For this reason, many methods have been developed to estimate species trees that can be accurate even with high levels of ILS (*2, 6, 9–16*). Some of these methods co-estimate gene trees and species trees directly from sequence alignments (e.g., BEST (*11, 12*) and *BEAST (*15*)), and others, called "summary methods", operate by combining estimated gene trees into a species tree. Some of these summary methods (e.g., MP-EST (*10*) and the population tree from BUCKy (*9*)) are based on the multi-species coalescent model (*17*), and have been proven to be statistically consistent under that model, which means they will reconstruct the true species tree with high probability given a sufficiently large number of true gene trees (*9, 10, 14, 15*). Together, these studies suggest that statistically consistent coalescent-based methods should have better accuracy than concatenation (which is conjectured to not be statistically consistent in the presence of ILS) (*8, 18*).

Nevertheless, the performance of coalescent-based methods has been mixed. Co-estimation methods have excellent accuracy but typically use Bayesian MCMC techniques that do not converge quickly on datasets with hundreds of genes (e.g., the 5-species 166-locus dataset of (*19*) required 1 billion MCMC iterations to reach convergence using *BEAST); hence, only summary methods are feasible for large datasets. Some summary methods can give good results on some biological datasets (*11, 12*), but on other datasets the summary methods have not been able to produce well-resolved and highly supported trees (*20*), even with a large quantity of data (*21*). While high branch support does not imply high accuracy, simulation studies have also shown that summary methods are not always as accurate as concatenation (*22, 23*), in that species trees estimated using summary methods can have higher missing branch rates than concatenation even in the presence of substantial ILS. While there may be several reasons for this disparity in performance, "poor phylogenetic signal" in individual genes is a potential problem for coalescent-based summary methods (*23,24*). Furthermore, the statistical guarantees of summary methods have been proven only for the case where the input is a sufficiently large number of *true gene trees* sampled from the distribution defined by the species tree. Biologically this means that statistically consistent methods have performance guarantees given a sufficiently large number of unlinked gene trees for which the only cause of incongruence with the species tree is ILS, and which are estimated without any error. While the first requirements may be satisfied in some cases, the last requirement of error-free estimation is much harder, since many conditions (including short edges in gene trees) make completely accurate gene tree estimation highly unlikely. Thus, the theoretical guarantees of coalescent-based methods may not hold under conditions in which estimated gene trees have even a small amount of error.

Genome-scale phylogenetic analyses can utilize very large numbers of genomic markers to estimate the species tree; however, when the markers have reduced phylogenetic signal, the accuracy of summary methods is also reduced (*23*). Therefore, while species tree estimation in the presence of ILS is inherently challenging, it is especially so when there are markers with reduced phylogenetic signal, a condition that is highly likely to occur for rapid radiations.

This was the challenge that confronted the avian phylogenomics project (*25*). The species tree estimated using the concatenated maximum likelihood analysis had a succession of short branches suggestive of a radiation (which increases the probability of ILS), and there was also evidence of massive gene tree conflict (e.g., the average topological distance between estimated gene trees and the concatenated analysis was 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% (trees based on exons and UCEs having average BS below 40%), and less than 4% of the gene trees had average BS above 60% (Fig. S1). Therefore, the large topological distances between estimated trees are to some extent the result of poor phylogenetic signal in the gene sequences; however, evidence of strongly supported gene tree conflict remained even after taking these low bootstrap support values into account (SOM (*26*) Section 2.1.1 and Fig. S12). Thus, not only is there substantial evidence of ILS (reducing the accuracy of concatenation), but the gene trees are generally poorly estimated (reducing the accuracy of summary methods).

Constructing phylogenies from genes with low phylogenetic signal is challenging, even if ILS is not an issue, and several approaches for selecting markers for use within a concatenated analysis have been suggested (*27*). However, restricting markers to a subset of “good” data presents a potential problem to statistically-based summary methods, because statistical guarantees depend on having a large enough random sample of true gene trees. Therefore, removing genes (even if they have low signal) has the potential to bias the analysis, since it uses a biased sample of the distribution.

In this paper, we present a new "statistical binning" technique that enables coalescent-based summary methods to obtain high accuracy without necessitating the restriction of the data to a subset of the markers. We show that using this technique with MP-EST (*10*), a leading summary method, produces dramatically improved accuracy under biologically realistic conditions. This binned version of MP-EST can achieve greater accuracy than concatenation even when unbinned MP-EST is less accurate than concatenation. We use this technique to analyze five biological datasets, including the mammalian dataset of (*4*), and to construct the first genome-scale coalescent-based estimation of the avian tree of life (*25*).

Statistical Binning Technique

The "statistical binning" approach we have developed to improve coalescent-based summary methods (Fig. 1) combines statistical and graph optimization techniques. We partition the gene set into bins, so that each bin defines a set of genes that passes a statistical test for "combinability" and the bins have approximately the same size. After the genes are partitioned into bins, we concatenate the alignments of the genes into one large "supergene alignment", and we compute trees on each supergene alignment using maximum likelihood (*28*); this produces a set of "supergene trees" (one for each bin). We then construct a species tree from the set of supergene trees using the preferred summary method. Thus, each summary method can be used with this pipeline to produce a "binned" version of the method, and the original version of the method (in which each gene tree is in its own bin) is the "unbinned" version of the method. See Methods (below) for more details on the binning technique.

Evaluation

We use biological and simulated datasets to evaluate species trees estimated using binned and unbinned summary methods, as well as concatenation using maximum likelihood under the

GTRGAMMA model, computed by RAxML (*28*). We study three summary methods: the greedy consensus (also known as the extended majority consensus), Matrix Representation with Parsimony (MRP) (*29*), and MP-EST (*10*) (SOM Section 3.7). We chose MP-EST because it has been proven statistically consistent under the multi-species coalescence model, has been used in many recent systematic studies (*4, 30–32*), and previous studies have suggested it has better performance than other summary methods (*10*). The greedy consensus is not statistically consistent (*33*), but it is not known if MRP and concatenation are, and some simulation studies (e.g., (*8*)) suggest that concatenation is not.

For the simulated datasets, we use two model species trees: one based on the avian dataset with 48 species and 14k loci studied by Jarvis *et al.* (*25*), and one based on a mammalian dataset with 37 species and 447 loci studied in (*4*) (see SOM Section 3.2 for details). Each model species tree was computed using MP-EST from the biological data, and we simulated gene trees within the species trees (Table S1) based on the multi-species coalescent model (*17*). We modified the branch lengths on the model species trees to vary the amount of ILS, either doubling branch lengths (and reducing ILS) or halving the branch lengths (and increasing ILS).

The avian biological gene trees have very low average BS: of the three types of genomic markers (exons, introns, and UCEs) used in that analysis, the exons had the least signal (average BS 24%), the introns had the most (average BS 48%), and the UCEs were intermediate in support (average BS 39%). The longest introns (with at least 10,000 bp) had the highest average BS (59%), but represented a very small fraction of the total set of gene trees (only 638 of more than 14,000). We created model conditions that resembled these four types of markers (exons-only, UCEs-only, introns-only, and long introns-only) with respect to their average BS values, and refer to these different model conditions by the partition type. The simulated mammalian datasets have two support levels (63% and 79%), bracketing the average BS values in the biological data (71%). We varied the number of genes from 200 to 2000 for the avian dataset, and from 200 to 800 for the mammalian dataset. Finally, we created mixed model conditions, one with 14,350 genes for the avian simulation experiment and the other with 400 genes for the mammalian simulation experiment, to closely approximate the biological datasets in terms of the number of loci and average BS.

Accurate species tree estimation for the avian simulated datasets is particularly challenging because of the very high ILS level and very low BS values in most of the partitions. The mammalian simulated datasets have lower ILS levels and higher BS values, and so are easier to analyze well. Thus the simulation study allows us to explore performance under a range of relatively easy (mammalian data) to extremely challenging (avian data) conditions.

For the simulated datasets, we measure topological error using the missing branch (false negative) and false positive rates for both estimated gene trees and species trees. We measure how well the distribution on gene trees is estimated by comparing triplet frequency distributions calculated based on true gene trees and estimated gene trees. We measure the estimation error in the branch lengths estimated by MP-EST (which measure the ILS in the data). See SOM Section 3.6 for a detailed discussion of error measurement.

Results

The default model conditions are centered on the properties of the avian and mammalian datasets in terms of average gene tree bootstrap support and ILS levels, but we vary these defaults to include lower and higher ILS levels, and estimated gene trees with very low or very high BS, in order to understand the impact of binning under a wider range of conditions. We present results for unbinned MP-EST, binned MP-EST, and concatenation; results for MRP and Greedy are shown in the SOM (Fig. S6 and Fig. S8). See the SOM Section 2.2 for a deeper discussion.

Figure 2A shows results using 1000 avian gene trees, with varying gene tree support (Table S2 provides standard deviations). Binned MP-EST is consistently and significantly more accurate than concatenation ($p < 10^{-5}$, two-way ANOVA test with Benjamini-Hochberg (BH) correction, Table S4), and is also significantly better than unbinned MP-EST ($p < 10^{-5}$). Improvements obtained by binning over unbinned MP-EST are impacted by gene tree support ($p = 0.003$ for the interaction effect; Table S5). The largest improvement is for the exon-like genes, where using binning reduces the error by 42% (from 24% to 14%). The gap between binned MP-EST and unbinned MP-EST decreases with the increase in gene tree BS values, but only for the highest BS genes (long intron-like) do binned and unbinned MP-EST have roughly the same accuracy. Concatenation is generally more accurate than unbinned MP-EST, except for the highest support genes.

The results of varying the number of genes while fixing the support to either UCE-like (moderate BS) or intron-like (high BS) gene trees (Figure 2B and C) show that binned MP-EST is more accurate than unbinned MP-EST ($p < 10^{-5}$ for UCE-like and $p = 0.015$ for intron-like). Furthermore, the gap between unbinned and binned MP-EST grows with the number of genes. For example, with 2000 UCE-like genes, using binning reduces the error by 60%, from 16.4% to 6.7%. Binned MP-EST tends to be more accurate than concatenation, but the differences are not statistically significant. However, the overall differences between binned MP-EST and concatenation are close to significant for intron-like genes ($p = 0.076$). The improvement of binned MP-EST over concatenation increases with the number of genes; for example, with 2000 intron-like genes, the error for MP-EST is 3.3%, but concatenation has 8.4% error, and the differences are statistically significant ($p = 0.004$, paired one-sided Wilcoxon test). Finally, on the mixed model condition, binned MP-EST and concatenation each had 7% error, while all the other methods had at least 11% error (Fig. S6).

On the simulated mammalian datasets, for every number of genes and both ILS levels, binned MP-EST either matches or improves upon unbinned MP-EST and is more accurate than concatenation (Fig. 3A). On the moderate (63%) BS genes, binned MP-EST has the best accuracy for all numbers of genes we tested, followed closely by concatenation (differences not statistically significant), then by unbinned MP-EST (a significant improvement, $p = 0.001$). Improvements of binned MP-EST over unbinned MP-EST can be dramatic: for example, for 800 genes, binned MP-EST has 1.8% error and unbinned MP-EST has 5.9% error; thus, using binning reduces error by 69%. The results on the higher BS (79%) genes show slightly different results – binned MP-EST has the best accuracy at 200 genes, then binned and unbinned MP-EST have the same accuracy at 400 and 800 genes (overall differences are not statistically significant), and concatenation is the least accurate method and significantly worse than binned MP-EST ($p = 0.008$). Finally, on the mixed model condition that resembles the real mammalian dataset in terms of the number of genes and gene tree support, binned MP-EST has only 1.8% error, concatenation has 3.7% error, and unbinned MP-EST has 4.6% error (Fig. 3B).

Statistical binning has a very large impact on species tree branch length estimation as well: MP-EST always underestimates species tree branch lengths (in coalescent units) when analyzing estimated gene trees (in some cases by close to an order of magnitude), whereas the binned MP-EST trees have much more accurate branch lengths (Fig. 2D, Fig. 2E, Fig. 2F, and Fig. 3C). Because branch lengths are model parameters that determine the amount of ILS, underestimating branch lengths directly means overestimating ILS.

We explore the impact of the amount of ILS by scaling the species tree branch lengths. Binned MP-EST has better average performance than both unbinned MP-EST and concatenation regardless of the amount of ILS (Fig. 4). The differences between binned and unbinned MP-EST are statistically significant for both avian and mammalian datasets ($p < 10^{-5}$ and $p = 0.0002$ respectively), and differences between binned MP-EST and concatenation are significant on the avian dataset ($p = 0.025$). Furthermore, on the avian dataset, reducing the ILS level (2X condition) increases the impact of binning, and increasing the ILS level (0.5X condition) decreases the impact ($p = 0.0004$ for the interaction effect). The impact of ILS level on the mammalian dataset is similar, but less pronounced and not statistically significant. For the reduced ILS (2X) models on both the avian and mammalian datasets, binned MP-EST is dramatically more accurate than unbinned MP-EST at estimating species tree topologies and branch lengths. For example, with 1000 UCE-like avian genes, unbinned MP-EST has 16.2% tree error while binned MP-EST has only 4.9%, a reduction of 70%. In fact, on the reduced ILS avian and mammalian datasets, unbinned MP-EST is less accurate than concatenation and binned MP-EST is more accurate.

Performance on true gene trees is clearly an upper bound on what a summary method can achieve on estimated trees. MP-EST has outstanding accuracy on true gene trees, both for topology and branch length estimation, but is much less accurate on estimated gene trees (Figs. 2-4). We therefore explored the possibility that the improvement seen in species tree topology and branch length estimation is the direct result of an improvement in gene tree estimation accuracy (see SOM 3.3 for details).

Statistical binning improves the estimation of gene trees and triplet gene tree distributions, often dramatically (Table 1, but see also Figures S2 and S3). Reductions in gene tree estimation error are largest for the exon-like genes and decrease with the gene tree support. The reductions in triplet gene tree distribution estimation error are even larger (Table 1, Fig. S4 and Fig. S5), and the largest impact is for genes with the lowest support. The drops in gene tree distribution error are important because MP-EST uses estimated gene tree distributions to construct a species tree. Hence, improving the gene tree distribution estimation improves MP-EST's ability to estimate the species tree topology and branch lengths. The summary methods Greedy and MRP are defined by bipartitions in estimated gene trees, and so are also impacted by gene tree estimation error, a trend seen in Figures S6 and S8.

Biological Datasets

We studied five biological datasets: the avian dataset (*25*) with 14k genes and 48 species, the mammalian dataset studied by Song *et al.* (*4*) with 447 genes and 37 species, and three datasets studied by Salichos and Rokas (*27*): yeasts with 23 species and 1070 genes, vertebrates with 15 species and 1087 genes, and metazoa with 21 species and 225 genes. Each of these datasets shows evidence of gene tree discord (Fig. S13), but vary with respect to the average BS (Fig. S14): 32% for avian, 49% for metzoa, 71% for mammals, 72% for yeasts, and 76% for vertebrates.

Species trees estimated using concatenation and gene trees were available from the respective publications, except for the gene trees from the mammalian dataset which we recomputed. We use partitioning based on genes when estimating supergene trees for the avian, metazoa, vertebrates, and yeast datasets (SOM Section 3.4). Results on the avian data are reported in detail in (*25*) and are briefly summarized here. See Table 2 for a summary of the results obtained using binned and unbinned MP-EST on these data, and Figure S10 for bin sizes.

Avian. The avian dataset represents the most challenging of the biological datasets we studied - very low average bootstrap support for almost all markers (Fig. S1) and large topological distances between estimated gene trees. On the avian dataset, an unbinned MP-EST analysis of the full 14K loci produced a tree (Fig. 5A) that has low to moderate support for some branches and fails to recover four key clades (Columbea [flamingo, grebe, pigeon, mesite, sandgrouse], Cursores [crane, killdeer], Otidimorphae [bustard, turaco, cuckoo], and Psittacopasseres [parrot, passerine, falcon, seriema]) that are recovered consistently with other types of reliable analyses on the avian full genome dataset, as reported in (*25*). Failure to recover Psittacopasseres is particularly surprising, as it has been recovered across different studies and types of data (*20, 21, 34, 35*). In contrast, binned MP-EST on all 14K loci (Fig. 5(A)) has more edges with high support and recovers all key clades that are recovered by other analysis of the same data in (*25*); this tree is presented as one of the two main hypotheses in (*25*), along with a partitioned concatenated analysis using maximum likelihood (Fig. S17).

An unbinned MP-EST tree based on the introns-only dataset (*25*) produced a fairly well supported MP-EST tree (31 out of 45 edges have 100% support and 34 have 95% or higher) and recovered all the key clades missing from the unbinned MP-EST tree computed on the full set of 14K loci. However, the binned MP-EST analysis (Fig. S16) on the introns-only dataset also recovers all the key clades, and has even higher support (33 edges with 100% support and 35 with 95% support or more), with increased support for some key novel clades (e.g., from 82% to 100% for Columbea).

Metazoa. The Metazoan dataset also represents a challenging analysis, since the average bootstrap support was also very low (only 49%). The most important difference between the unbinned and binned MP-EST trees (Fig. 5B) is among Chordates, where the unbinned MP-EST tree puts Cephalochordates (represented by *B. floridae*) as sister to vertebrates (Craniates), and the binned MP-EST tree puts Urochordates (represented by *C. intestinalis*) as sister to vertebrates. While Cephalochordates were traditionally thought to be the sister to all the extant vertebrates (*36*), the majority of the recent molecular evidence strongly supports Urochordates as the sister to all vertebrates (*37–39*), and hence the binned MP-EST is likely correct. There are also some differences between the two trees within Protostomia, but here both MP-EST trees have low support and neither is congruent with the literature (see SOM Section 2.1.3 for more discussion).

Mammalian. The mammalian dataset has gene trees with substantially higher average BS (71%), but also demonstrates substantial gene tree incongruence suggestive of ILS. Song *et al.* (*4*) found some differences between MP-EST and concatenation regarding the placement of tree shrews and bats: Scandentia (tree shrews) is sister to Glires (Rodentia/Lagomorpha) in their concatenated analysis, but is sister to primates in their MP-EST tree. We re-analyzed this dataset and identified 21 genes with mis-labeled sequences (subsequently confirmed (*40*)) plus two outlier genes (Fig. S15 and SOM Section 2.1.2). We removed these 23 genes, and re-analyzed the data using concatenation and both binned and unbinned MP-EST (Fig. S18).

Our concatenation tree is topologically identical to the concatenation tree in (*4*). The unbinned MP-EST tree on this reduced gene set is similar to the unbinned MP-EST tree reported in (*4*), but has lower support for tree shrews being sister to primates (99% in (*4*), 64% with our analysis), and one topological difference among low support edges. We confirmed these differences are not related to the filtering of those 23 genes; instead, the likely source of this difference is that Song *et al.* resampled both genes and sites in their bootstrapping procedure, but we resampled only sites.

The binned MP-EST and unbinned MP-EST trees on the reduced gene set are very similar, but have one notable difference: in the binned MP-EST tree, tree shrews are sister to Glires with 80% support, just like their position in the concatenation tree. Thus, the placement of Scandentia, and whether it is sister to primates or to Glires, depends on the mode of analysis. This agreement between the binned MP-EST analysis and concatenated analysis of the reduced dataset represents a potentially important finding, but contradicts Janecka *et al.* (*41*) (which specifically addressed this question) and Boussau *et al.* (*42*). However, these two analyses did not take ILS into account, and it is possible that ILS is at the source of the difficult placement of this species.

The binned MP-EST tree still differs from the concatenation tree with respect to the location of Chiroptera (bats), which are sister to all other Laurasiatheria except the basal Eulipotyphla in the binned MP-EST tree, just as in the unbinned MP-EST tree.

Yeast. The yeast dataset has relatively high average BS (72%). The binned and unbinned MP-EST topologies are identical, and both have 100% support for all but one branch (Fig. S21). See SOM Section 2.1.5 for more discussion.

Vertebrates. The vertebrate dataset had the highest average BS (76%) of all datasets we examined. Both binned and unbinned MP-EST trees had the same topology, and both are topologically identical to the concatenation tree on the same data (Fig. S20). See SOM section 2.1.4 for more discussion.

Discussion

Performance on simulated data demonstrates the benefits of binning under a wide range of model conditions, Binning improves many aspects of coalescent-based estimation, including species tree topology and branch lengths, gene tree topology, and gene tree distribution. Consequently, unbinned MP-EST over-estimates ILS levels on the simulated datasets, while binned MP-EST comes much closer to the correct ILS levels. All these improvements result from improvements in gene tree topology and gene tree distribution estimation. Binning produces very substantial reductions in error under conditions that were found in the avian and mammalian datasets – conditions that are likely to be common in phylogenomic analyses for species trees with ILS, as these involve large numbers of genes whose trees have short branches, and hence many estimated gene trees with at best moderate bootstrap support. Also, although unbinned methods are rarely more accurate than concatenation, binned MP-EST is almost always *at least as* accurate as concatenation, and there are many model conditions in which binned MP-EST is more accurate than concatenation and unbinned MP-EST is less accurate than concatenation.

The five biological datasets provide important insights into the impact of binning under specific biologically realistic conditions. As expected, binning has a large impact on datasets with less well supported gene trees (avian and metazoan), and has little impact on the yeast and vertebrate datasets, both of which had very well resolved gene trees. Interestingly, binning does impact the MP-EST analysis of the mammalian dataset, which also had fairly well resolved gene trees. Where binning has an impact, binned MP-EST typically produces trees that are in closer agreement with accepted reconstructions than unbinned MP-EST. For example, the unbinned MP-EST analysis of the avian dataset fails to recover monophyly of several key clades, including Psittacopasseres (a clade that is reliably recovered by many phylogenetic analyses on many datasets), but binned MP-EST recovers all the key avian clades. Binning reduces gene tree incongruence on biological datasets (Fig. S13), suggesting that binned MP-EST may not overestimate ILS on biological datasets as much as unbinned MP-EST does. This trend is consistent with performance on simulated data, and suggests that more accurate estimates of ILS in biological data may be obtained through the use of statistical binning.

Binning has the potential to group genes together that have different true topologies, but this will only happen if the two gene trees have conflicting branches that have bootstrap support that is too low to prohibit combining the two genes, or when there is enough estimation error in the gene trees so that they do not display those pairs of conflicting branches. Thus, binning does not have theoretical statistical consistency guarantees, and can in theory result in incorrect estimation of gene tree distributions. However, we observe the opposite in our simulated studies – estimated gene tree distributions are *more accurate* after binning. We suggest the following explanation. Binning will never group genes with different topologies together unless the conflicting branches had low support; however, in the conditions we are considering, low support is typically a result of insufficient phylogenetic signal, and thus low support branches have a high probability of being incorrectly reconstructed. As we have seen, the inclusion of poorly estimated gene trees distorts the estimated triplet gene tree distribution. However, binning is able to improve the estimated triplet gene tree distribution by reducing the noise, suggesting that the beneficial impact of noise reduction produced by binning is stronger than the detrimental impact of potentially putting together genes that have different topologies. In other words, although there is a potential for genes with different trees to be put in the same bin, the overall impact of binning is beneficial rather than detrimental. Finally, when gene trees are accurately estimated, they also tend to have high support, and so if two highly accurately estimated gene trees have different topologies, they will not be binned together; we observe this in our simulation study, where bin sizes shrink as support increases (Table S6).

While gene tree estimation error can come from insufficient phylogenetic signal in the gene sequences, it can also come from poorly estimated alignments (*43*) or errors introduced during the tree inference due to sequences evolving according to complex processes (e.g., heterotachy and base composition variation) that violate the assumptions of the inference model (*44, 45*). As this study has shown, binning effectively reduces error due to insufficient phylogenetic signal, but we have no evidence that it could reduce phylogenetic error due to alignment error or model misspecification. Consequently, for optimal performance, appropriate care should be devoted to obtaining good alignments and choosing an adequate model of sequence evolution to reconstruct both gene and supergene trees. The construction of supergene trees in particular could be performed using models partitioned according to the genes that make up the supergene, as was done here for the biological datasets.

Statistical binning is a technique to improve gene tree estimation; therefore, there are limits to the conditions under which statistical binning can provide any benefit. For example, if all estimated gene trees have close to 100% bootstrap support, then binning will not group genes together except when they share the same topology. At the other extreme, if all gene trees have extremely low support (lower than the exon-like genes in the avian simulation), then binning will approach (or be equal to) a completely concatenated analysis. Thus, binning will have little impact with truly excellent data, and will reduce to concatenation with extremely poor data.

The first case (near perfect data) is not a problem, because the best summary methods (e.g., MP-EST) have excellent accuracy on highly accurate gene trees. The second case, however, may be inappropriate for any summary method – there may be too much noise in each gene tree for a summary method to perform well. However, this study showed that statistical binning is helpful for a wide range of simulation conditions, including very poorly estimated gene trees (the exon-like avian genes with 25% average bootstrap support) and very accurately estimated gene trees (e.g., 200 mammalian genes with 79% bootstrap support). Statistical binning also improved almost all the biological dataset analyses, leaving only the yeast and vertebrate analyses unaffected. Thus, while statistical binning has limits, in these analyses it never reduced accuracy and almost always improved accuracy - often very substantially.

Statistical binning is not a coalescent-based species tree estimation method *per se*, but rather a method that improves gene tree estimation; therefore, it is a general technique for improving coalescent-based summary methods. Furthermore, statistical binning can be used to improve summary methods for species tree estimation when gene tree discord results from other biological phenomena, including horizontal gene transfer or hybrid speciation. Supergenes produced by our method correspond to groups of genes sharing the same or similar tree topologies; in some cases, such groups of genes could be syntenic or linked, but they do not have to be. Indeed, our method is agnostic as to the technique used to predict that two or more genes share the same topology. As a consequence, binning also could be used for purposes other than finding the species tree, for example, as a way to detect outlier genes, or genes that have been co-transferred. Depending on the objective, introducing additional information such as chromosomal position into the binning algorithm could further improve the definition of supergenes.

Because the only source of discord between true gene trees and true species trees in our simulation is ILS, and not also horizontal gene transfer, gene flow, recombination, gene duplication and loss, or model misspecification, the model conditions favor MP-EST (which is based on the same model used for simulations) over concatenation (which assumes no ILS is present). Given this, the fact that unbinned MP-EST is less accurate than concatenation in many conditions is noteworthy. However, a more extensive analysis under a wider range of model conditions in which other sources of gene tree discord were included would enable a better understanding of the relative accuracy of concatenation and coalescent-based species tree estimation.

Summary

Our study provides evidence on both simulated and biological data that coalescent-based summary methods for estimating species trees can have poor accuracy in the presence of even moderate amounts of gene tree estimation error, thus presenting substantial challenges for genome-scale species tree estimation. The Jarvis *et al.* (*25*) avian dataset of roughly 14,000 markers with mostly very low phylogenetic signal and patterns of massive gene tree incongruence consistent with ILS presented a prime example of a difficult species tree problem with these properties, for which a coalescent-based analysis (using MP-EST, a leading coalescent-based method) conflicted substantially with the best maximum likelihood tree obtained on the dataset.

The statistical binning technique we present in this paper is designed to enable highly accurate species tree estimations in the presence of poorly estimated gene trees; however, the study shows that it can also improve species tree estimation more generally. Our study shows that statistical binning reduces gene tree estimation error, so that summary methods (such as MP-EST) have improved species tree topology and branch length estimation. The study also shows that binned MP-EST has better accuracy than unbinned MP-EST on many biological datasets and under a wide range of realistic model conditions, and is almost always more accurate than concatenation, even for cases where unbinned MP-EST is less accurate than concatenation.

Based on the improved accuracy of binned MP-EST over unbinned MP-EST, we used binned MP-EST to analyze the avian phylogenomic dataset, obtaining a tree that was highly congruent with the concatenation analysis. This tree is one of the two main phylogenetic hypotheses presented in the avian phylogenomics paper (*25*). Thus, statistical binning enabled the first coalescent-based estimation of the avian Tree of Life at a genome-scale.

Methods

The statistical binning technique includes a combinability test that evaluates whether a given pair of genes are likely to have significant topological incongruence, so that a partitioned concatenated analysis of those two genes is likely to be problematic. Because supergene trees can be estimated using partitioned concatenated analyses (which would allow the branch lengths and other GTR parameters to be re-estimated for each gene within a partition), we need only consider topological incongruence, allowing genes whose estimated trees share the same topology but different branch lengths to be placed in the same bin.

We use maximum likelihood with bootstrapping to estimate gene trees with branch support values, and we say that a given pair of trees "exhibit conflict at threshold t" if there is a pair of "incompatible branches", one in each of the two gene trees, each with bootstrap support of at least t. Two trees that do not exhibit conflict at threshold t are "combinable", and a set of trees for which all pairs are combinable is a "combinable set".

Saying that two branches are incompatible means that there is no tree that has both of these branches (*46*) (more specifically, no tree exists with branches that induce the bipartitions defined by these two branches). Thus, to test two trees for incompatibility at threshold t or higher, we collapse all branches in each tree with support below t, and then ask whether a tree exists that is a common refinement of these two collapsed trees. Testing for compatibility of two trees can be performed in polynomial time (*47*); hence, this calculation is fast. Alternative tests for combinability could also be used, with potentially improved results, but at a cost of increased running time; see (*48*) for a discussion of current tests for phylogenetic combinability.

The partitioning step uses a graph-based optimization, in which we build a graph in which each gene is represented by a node and an edge is present between two nodes (i.e., genes) if the estimated trees on that pair of genes exhibit conflict at threshold t. By definition, the graph depends on the parameter t; thus, smaller values for t will generally consider trees less likely to be combinable than larger values. If $t > 100\%$ were permitted, this would create a graph containing no edges, and all pairs of genes would be considered "combinable", and statistical binning would result in a concatenated analysis. However, binning genes into small subsets without any consideration for combinability (*23*) can also be employed.

The graph created is called an "incompatibility graph". To create bins from this graph, we color the vertices of the graph so that no two vertices with the same color are adjacent, and put all vertices with the same color into a common bin. Thus each bin consists of a combinable set (so that all pairwise incompatibility has support below t). Thus, adjusting the support threshold t enables more aggressive or conservative binning. Once bins are formed, alignments of genes in the same bin are concatenated into a supergene alignment, and supergene trees are estimated on these alignments using maximum likelihood. These supergene trees are used as input to the summary method of choice.

The input to the species tree estimation method is greatly influenced by the coloring step. Since the statistically consistent methods use the distribution of gene trees to estimate the species tree, it is important for the supergene tree distribution to be close to the gene tree distribution; for this reason, we seek a vertex coloring in which the different color classes have approximately the same size; we call this a "balanced coloring". We also seek a vertex coloring with a small number of colors, so that we have the largest bins we can, given the constraints imposed by combinability and balanced color class sizes. We call such a coloring a "balanced minimum vertex coloring". However, finding a balanced minimum vertex coloring is a constrained version of the NP-hard vertex coloring problem (*49*), and so it is unlikely that any exact algorithm for this problem will run in polynomial time. Therefore, we developed a heuristic for balanced minimum vertex coloring and used it to perform these analyses (SOM Section 3).

We set the statistical support threshold t as follows. We note that using 75% for the bootstrap support has been a standard threshold for branch reliability (*50*), and so 75% represents a reasonable setting for t; however, when the datasets are large, we can afford to be more conservative and pick a smaller threshold. We also explored the effect of the support threshold (Fig. S9) and saw that setting t to either 50% or 75% gave good results. Therefore, we set two thresholds: a conservative threshold of $t = 50\%$ that we use for datasets with at least 1000 genes, and a moderate threshold of $t = 75\%$ that we use for the other datasets.

References and Notes

1. A. Rokas, B. L. Williams, N. King, S. B. Carroll, *Nature* **425**, 798 (2003).

2. L. L. Knowles, *Systematic Biology* **58**, 463 (2009).

3. S. J. Hackett, *et al.*, *Science (New York, N.Y.)* **320**, 1763 (2008).

4. S. Song, L. Liu, S. V. Edwards, S. Wu, *Proceedings of the National Academy of Sciences of the United States of America* **109**, 14942 (2012).

5. R. W. Meredith, *et al.*, *Science (New York, N.Y.)* **334**, 521 (2011).

6. J. H. Degnan, N. A. Rosenberg, *Trends in Ecology & Evolution* **24**, 332 (2009).

7. S. V. Edwards, L. Liu, D. K. Pearl, *Proceedings of the National Academy of Sciences of the United States of America* **104**, 5936 (2007).

8. L. Kubatko, J. Degnan, *Systematic Biology* **56**, 17 (2007).

9. B. R. Larget, S. K. Kotha, C. N. Dewey, C. Ané, *Bioinformatics* **26**, 2910 (2010).

10. L. Liu, L. Yu, S. V. Edwards, *BMC Evolutionary Biology* **10**, 302 (2010).

11. L. Liu, D. K. Pearl, *Systematic Biology* **56**, 504 (2007).

12. L. Liu, *Bioinformatics* **24**, 2542 (2008).

13. L. Liu, L. Yu, D. K. Pearl, S. V. Edwards, *Systematic Biology* **58**, 468 (2009).

14. E. Mossel, S. Roch, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **7**, 166 (2010).

15. J. Heled, A. J. Drummond, *Molecular Biology and Evolution* **27**, 570 (2010).

16. Y. Yu, T. Warnow, L. Nakhleh, *Journal of Computational Biology* **18**, 1543 (2011).

17. B. Rannala, Z. Yang, *Genetics* **164**, 1645 (2003).

18. S. V. Edwards, *Evolution* **63**, 1 (2009).

19. B. T. Smith, M. G. Harvey, B. C. Faircloth, T. C. Glenn, R. T. Brumfield, *Systematic Biology* (2013).

20. R. T. Kimball, N. Wang, V. Heimer-McGinn, C. Ferguson, E. L. Braun, *Molecular Phylogenetics and Evolution* **69**, 1021 (2013).

21. J. E. McCormack, *et al.*, *PLoS ONE* **8**, e54848 (2013).

22. M. DeGiorgio, J. H. Degnan, *Molecular Biology and Evolution* **27**, 552 (2010).

23. M. S. Bayzid, T. Warnow, *Bioinformatics* **29**, 2277 (2013).

24. S. Patel, R. Kimball, E. Braun, *J Phylogen Evolution Biol* **1**, 2 (2013).

25. E. Jarvis, *et al.* (2013). Avian phylogenomics paper, in preparation for submission to Science.

26. S. Mirarab, M. S. Bayzid, B. Boussau, T. Warnow, Information on materials and methods is available on science online (2013).

27. L. Salichos, A. Rokas, *Nature* **497**, 327 (2013).

28. A. Stamatakis, *Bioinformatics* **22**, 2688 (2006).

29. M. A. Ragan, *Molecular Phylogenetics and Evolution* **1**, 53 (1992).

30. A. D. Leaché, R. B. Harris, B. Rannala, Z. Yang, *Systematic Biology* pp. 1–14, doi= 10.1093/sysbio/syt049 (2013).

31. L. Zhao, *et al.*, *PLoS ONE* **8**, e64642 (2013).

32. B. Zhong, L. Liu, Z. Yan, D. Penny, *Trends in Plant Science* **18**, 492 (2013).

33. J. Degnan, M. DeGiorgio, D. Bryant, N. Rosenberg, *Systematic Biology* **58**, 35 (2009).

34. A. Suh, *et al.*, *Nature Communications* **2** (2011).

35. N. Wang, E. Braun, R. Kimball, *Molecular Biology Evolution* **29**, 737 (2012).

36. C. Nielsen, *Animal evolution: interrelationships of the living phyla* (Oxford University Press, 2012).

37. F. Delsuc, H. Brinkmann, D. Chourrout, H. Philippe, *Nature* **439**, 965 (2006).

38. S. J. Bourlat, *et al.*, *Nature* **444**, 85 (2006).

39. T. R. Singh, *et al.*, *BMC Genomics* **10**, 534 (2009).

40. L. Liu, Personal communication to Tandy Warnow (2013). Email dated December 24, 2013.

41. J. E. Janecka, *et al.*, *Science (New York, N.Y.)* **318**, 792 (2007).

42. B. Boussau, G. Szöllsi, L. Duret, *Genome Research* **23**, 323 (2013).

43. K. Liu, S. Raghavan, S. Nelesen, C. R. Linder, T. Warnow, *Science* **324**, 1561 (2009).

44. J. Felsenstein, *Systematic Zoology* **27**, 401 (1978).

45. W. G. Weisburg, S. J. Giovannoni, C. R. Woese, *Systematic and Applied Microbiology* **11**, 128 (1989).

46. T. Warnow, *New Zealand Journal of Botany* **31**, 239 (1993).

47. T. Warnow, *Journal of Algorithms* **16**, 388 (1994).

48. J. Leigh, E. Susko, M. Baumgartner, A. Roger, *Systematic Biology* **57**, 104115 (2008).

49. E. Malaguti, P. Toth, *International Transactions in Operational Research* **17**, 1 (2010).

50. T. P. Wilcox, D. J. Zwickl, T. A. Heath, D. M. Hillis, *Molecular Phylogenetics and Evolution* **25**, 361 (2002).

51. SM was supported by a graduate fellowship from HHMI, MSB was supported by a graduate fellowship from the Fulbright Foundation, TW was supported by the U.S. National Science Foundation under grant DBI-1062335, and BB was supported by a postdoctoral fellowship from the Human Frontier Science Program, the CNRS, and Agence Nationale de la Recherche (ANR) through Grant ANR-10-BINF-01-01 Ancestrome. The authors thank Ed Braun, Scott Edwards, Tom Gilbert, Erich Jarvis, Liang Liu, David Posada, and Guojie Zhang for providing comments on earlier versions of the manuscript.

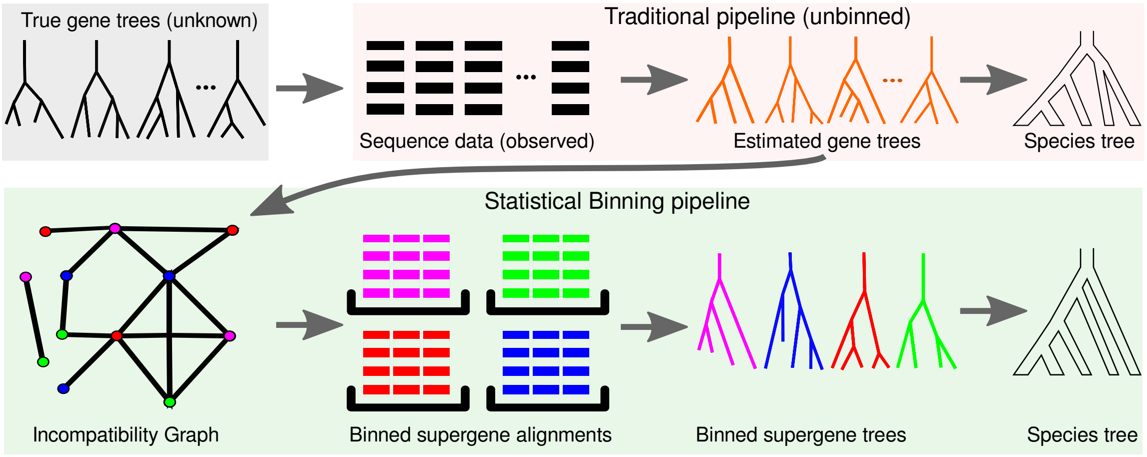

True gene trees (unknown)
Traditional pipeline (unbinned)
Sequence data (observed)
Estimated gene trees
Species tree
Statistical Binning pipeline
Incompatibility Graph
Binned supergene alignments
Binned supergene trees
Species tree

Figure 1: **Binning procedure**. Unknown true gene trees generate sequence data, from which estimates of gene trees can be obtained. In traditional pipelines, these estimated gene trees are used as input to species tree methods to generate a species tree. Statistical binning takes estimated gene trees and builds an incompatibility graph in which each node represents a gene tree and each edge represents a detected incompatibility between two gene trees at the given statistical support threshold t or higher. We use a heuristic we developed to color the nodes of the graph so that no two adjacent vertices have the same color, and so that the color classes are of similar sizes. This coloring of the vertices defines a partition of genes into bins (according to the color each is assigned) and ensures no two genes with statistically supported conflict are put in the same bin. For each bin, individual gene alignments are concatenated to get a "supergene" alignment, from which a supergene tree is estimated using maximum likelihood. The supergene trees are then used as input to the summary method of choice.

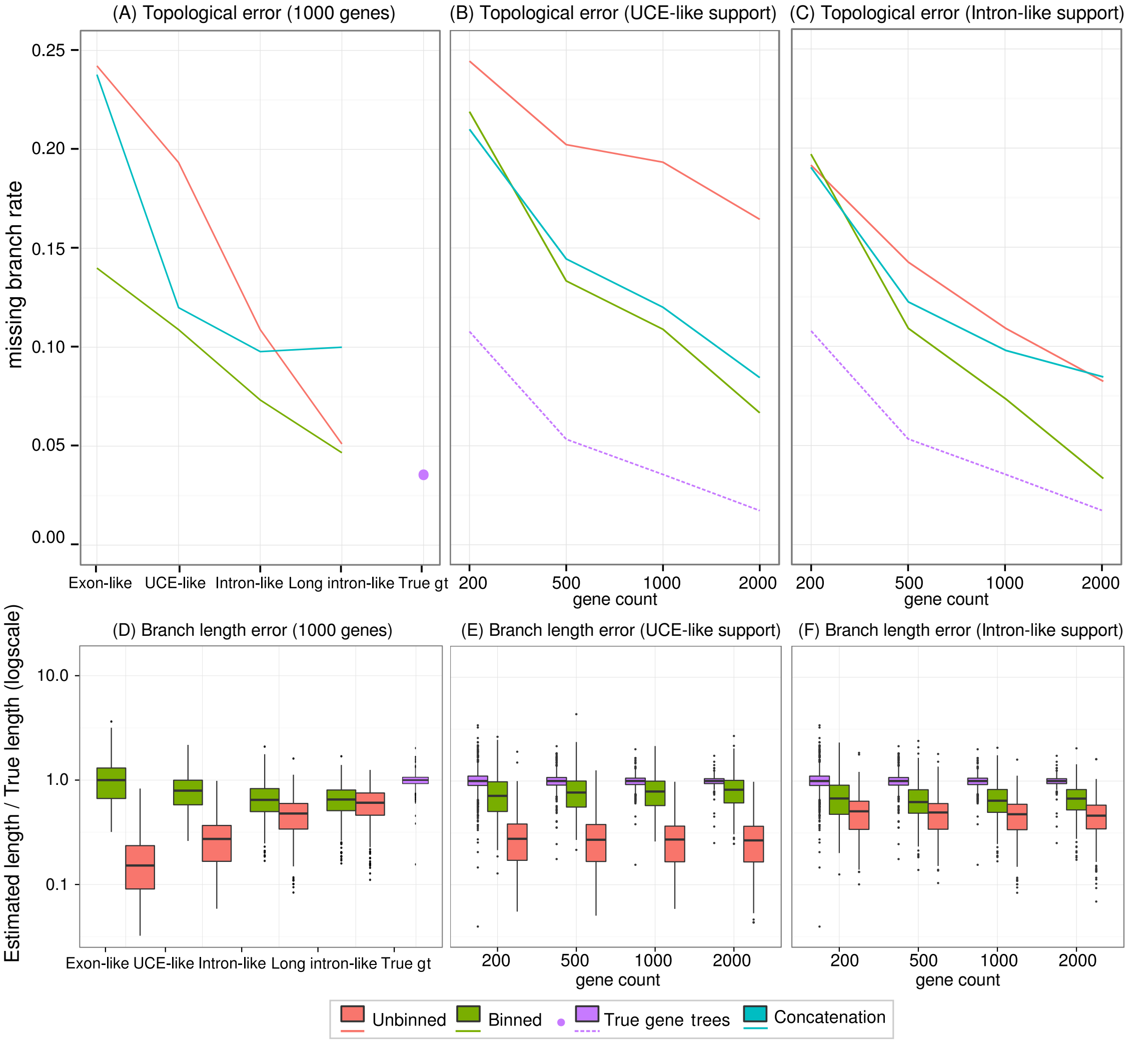


Figure 2: **Effect of binning on MP-EST on the simulated avian datasets.** Panels A-C show species tree topology error and panels D-F show species tree branch length error (boxplots with the ratio of estimated branch length to true branch length for branches of the true tree that appear in the estimated tree, 1 indicates correct estimation). MP-EST assigned an arbitrarily small length in the model tree to one branch, which we exclude from branch length calculations. Panels (A) and (D) fix the number of genes to 1000 genesand vary gene tree support. The other panels fix gene tree support to UCE-like (B and E) and intron-like (C and F) and vary the number of genes. Results are over 20 replicates for 200 genes conditions, and 10 replicates elsewhere. All gene trees are simulated under the 1X ILS condition. The improvement of binned MP-EST over unbinned MP-EST is statistically significant ($p < 10^{-5}$ for panels A and B, and $p = 0.015$ for panel C). The improvement of binned MP-EST over concatenation is statistically significant for panel A ($p < 10^{-5}$). See also Tables S2 (standard deviations), S4 (p-value), and S6 (bins sizes). See Figures S6 and S7 for results with MRP and Greedy.

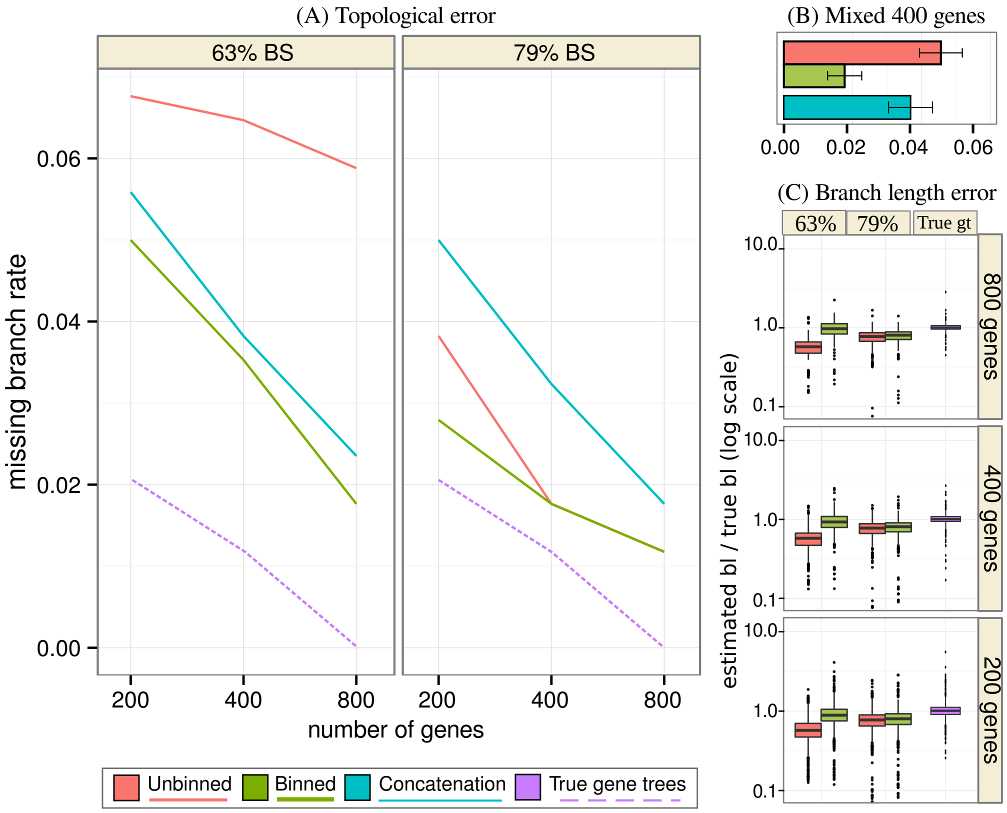


Figure 3: **Effect of binning on the simulated mammalian datasets.** Results are shown for concatenation and MP-EST applied to estimated gene trees with and without statistical binning, and also applied to the true gene trees. (A) Lines show average topological tree error (missing branch rate) over 20 replicates for 200 genes, 10 replicates for 400 genes, and 5 replicates for 800 genes. Results are shown separately for gene trees with 63% and 79% bootstrap support. (B) Topological error is shown for a mixed dataset with 200 genes of 63% BS level, and 200 genes of 79% BS level. (C) Error in branch lengths estimated by MP-EST in coalescent units is shown. The improvement in species tree topology produced by binned MP-EST over unbinned MP-EST is statistically significant ($p = 0.001$) for the 63% BS gene trees (panel A); there is also a statistically significant improvement of binned MP-EST over concatenation for the 79% BS gene trees (panel A, $p = 0.008$). See Tables S3 (standard deviation) and S4 (p-values). Results for MRP and Greedy are presented in Figure S8.

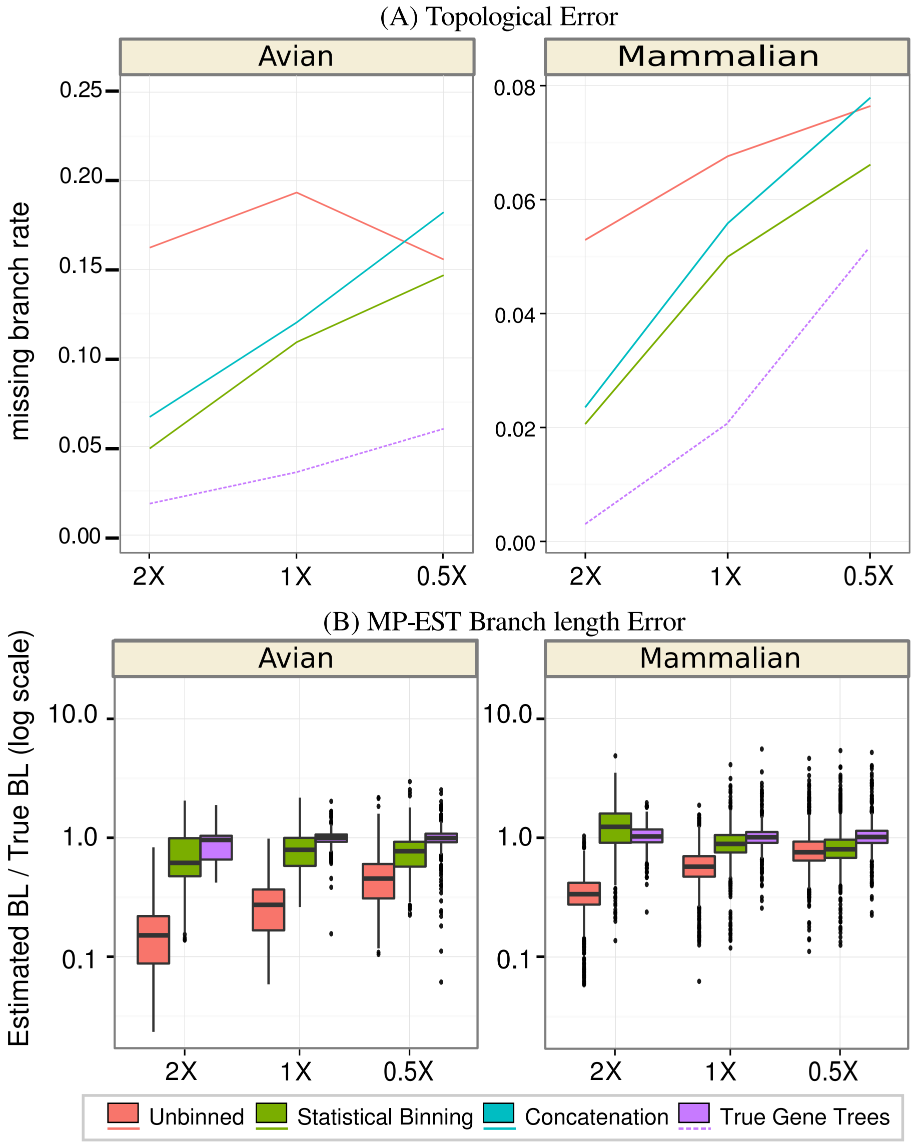


Figure 4: **Effects of ILS levels for the simulated avian and mammalian datasets.** Levels of ILS are changed by multiplying all branch lengths in the model species tree by a factor of 0.5 (to increase ILS) or 2 (to reduce ILS). (A) Topological tree error measured by average missing branch rate is shown. (B) Error in branch lengths estimated by MP-EST in coalescent units is shown. Results are over 10 replicates of 1000 UCE-like gene trees for avian, and 20 replicates of 200 gene trees with 63% BS for mammalian datasets Binned MP-EST has a statistically significant improvement over unbinned MP-EST ($p < 10^{-5}$ for panel A and $p = 0.00015$ for panel B). Binned MP-EST has a statistically significant improvement over concatenation ($p = 0.02504$) for panel A.

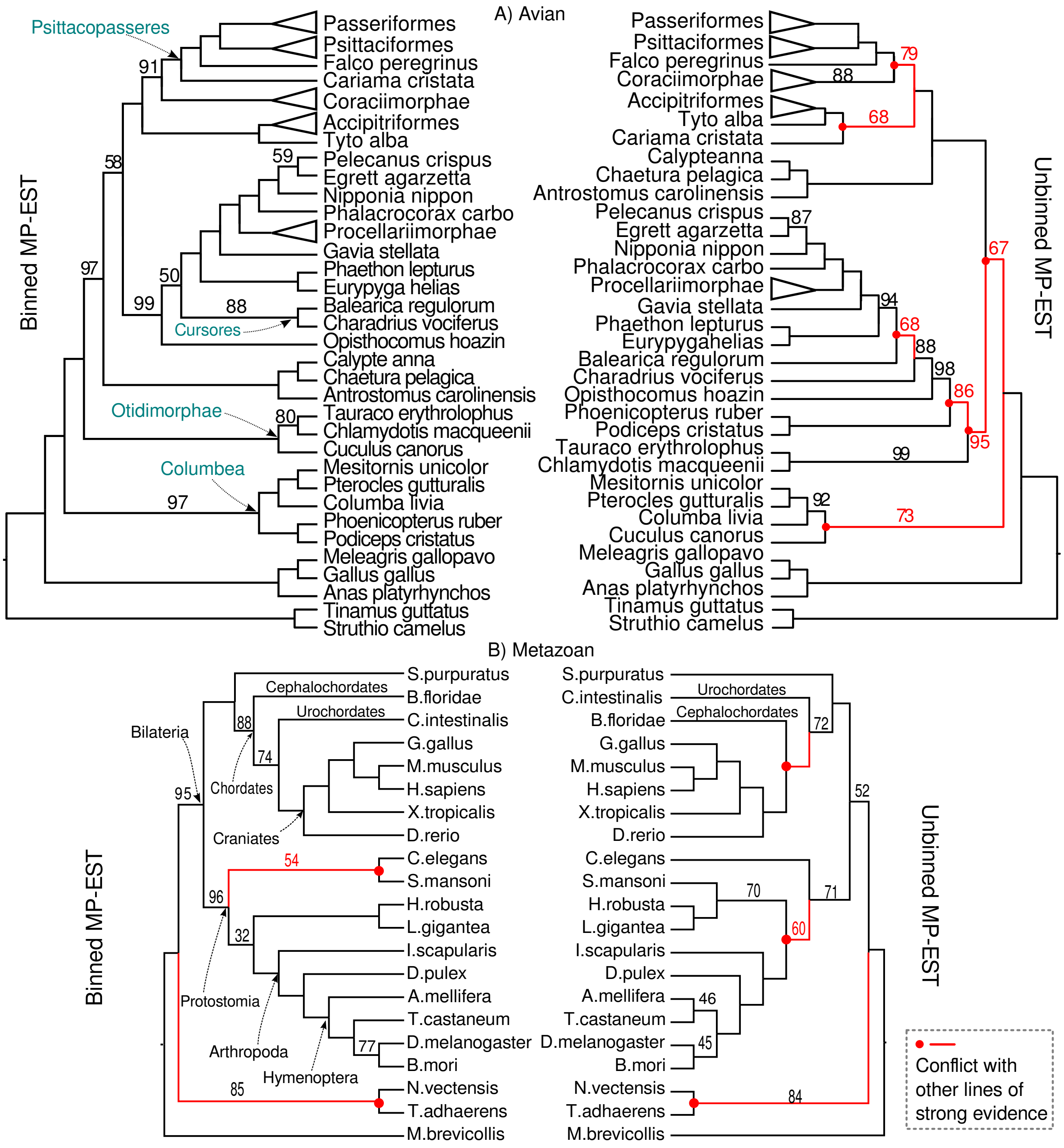
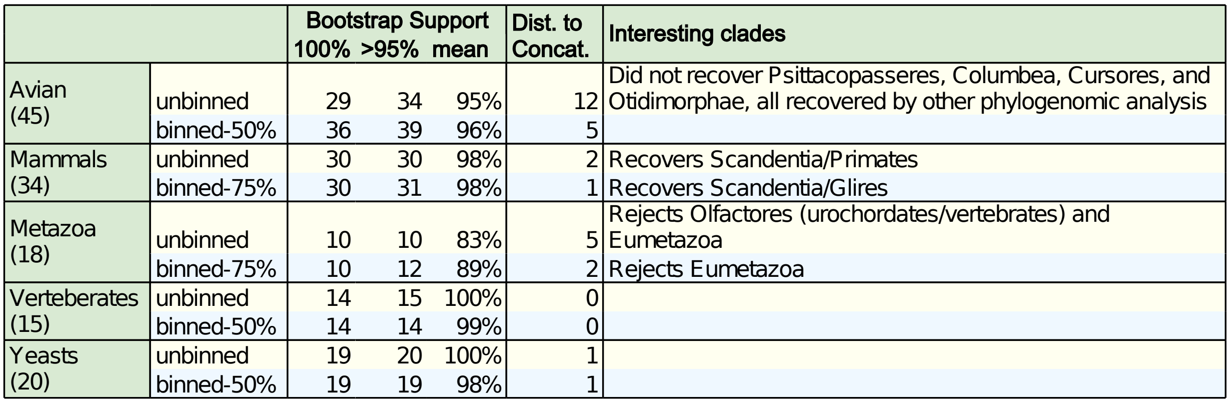


Figure 5: **Results on the (A) avian and (B) metazoan biological datasets using binned and unbinned MP-EST.** Bootstrap support is 100% on all branches, except those that show another value.

		Individual GT Error (RF)		GT Distribution Error (KL)	
		Unbinned	Binned	Unbinned	Binned
Avian	Exon-like (250bp)	79%	57%	0.234	0.025
(1000 genes)	UCE-like (500bp)	69%	57%	0.120	0.008
	Intron-like (1000bp)	55%	51%	0.033	0.008
	Long Int.-like (1500bp)	46%	45%	0.011	0.007
Mammalian	63% BS (500bp)	43%	35%	0.119	0.019
(200 genes)	79% BS (1000bp)	27%	26%	0.038	0.027

Table 1: **Gene tree estimation error, with and without binning for simulated datasets.** Results are shown for fixed number of genes (1000 for avian and 200 for mammalian) and levels of ILS (1X, i.e. observed), but also see Figures S2, S3, S4, and S5 for other conditions. Individual gene tree error is mean RF distance between true gene trees and all 200 bootstrap replicates of each estimated gene tree. For the supergene trees, each bootstrap replicate of each supergene tree is compared separately against each true gene tree for the genes put in that bin. We also characterize gene tree distributions by calculating the triplet frequencies for all $\binom{n}{3}$ possible triplets, and we do this both for true and estimated gene trees (using all 200 bootstrap replicates of all genes/supergenes in the case of estimated trees). Thus, we obtain a true and an estimated triplet frequency distribution for each of the $\binom{n}{3}$ triplets. We report the mean Kullback-Leibler (KL) divergence of the estimated distribution from the true distribution.



		Bootstrap Support			Dist. to Concat.	Interesting clades
		100%	>95%	mean		
Avian (45)	unbinned	29	34	95%	12	Did not recover Psittacopasseres, Columbea, Cursores, and Otidimorphae, all recovered by other phylogenomic analysis
	binned-50%	36	39	96%	5	
Mammals (34)	unbinned	30	30	98%	2	Recovers Scandentia/Primates
	binned-75%	30	31	98%	1	Recovers Scandentia/Glires
Metazoa (18)	unbinned	10	10	83%	5	Rejects Olfactores (urochordates/vertebrates) and Eumetazoa
	binned-75%	10	12	89%	2	Rejects Eumetazoa
Verteberates (15)	unbinned	14	15	100%	0	
	binned-50%	14	14	99%	0	
Yeasts (20)	unbinned	19	20	100%	1	
	binned-50%	19	19	98%	1	

Table 2: **Results on the biological datasets.** For the five biological datasets, MP-EST trees are compared in terms of support (the number of edges with support equal to 100%, edges with support at least 95%, and average support), the distance between concatenation and MP-EST trees (number of missing branches), and with respect to interesting differences between the two trees. The total number of branches in each tree is given parenthetically in the first column.

Supplementary Online Material

(Statistical binning improves species tree estimation in the presence of gene tree incongruence)

Siavash Mirarab,[1] Md Shamsuzzoha Bayzid,[1] Bastien Boussau,[2] Tandy Warnow[1*]

[1]Department of Computer Science, University of Texas at Austin,
[2]Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, France

[*]To whom correspondence should be addressed; E-mail: tandy@cs.utexas.edu

Contents

List of Figures

List of Tables

1 Supplementary Figures and Tables

1.1 Simulation properties

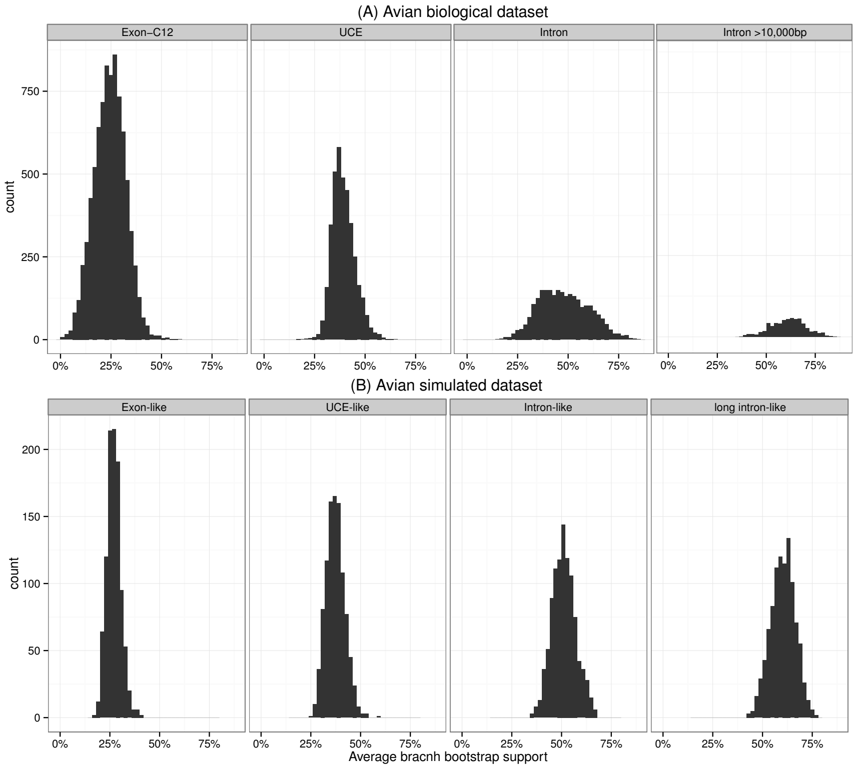


Figure S1: **Gene tree bootstrap support histograms for the avian simulated and biological datasets.** Histograms show the distribution of average bootstrap branch support across (A) three partitions of the avian dataset with a total of 14,446 loci (Jarvis *et al., submitted*), and (B) 1000 genes from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

	Mammals			Avian		
	2X	1X	0.5X	2X	1X	0.5X
Distance to Species Tree (mean)	18%	32%	54%	36%	47%	59%
Distance to Species Tree (min)	0%	3%	26%	16%	25%	36%
Distance to Species Tree (max)	42%	62%	82%	58%	67%	76%
Distance to other Gene Trees (mean)	26%	46%	71%	44%	57%	68%
Distance to other Gene Trees (min)	3%	14%	36%	20%	30%	44%
Distance to other Gene Trees (max)	51%	77%	96%	66%	75%	86%

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 10 replicates of 1000 genes for the avian dataset and 20 replicates of 200 genes for the mammals dataset.

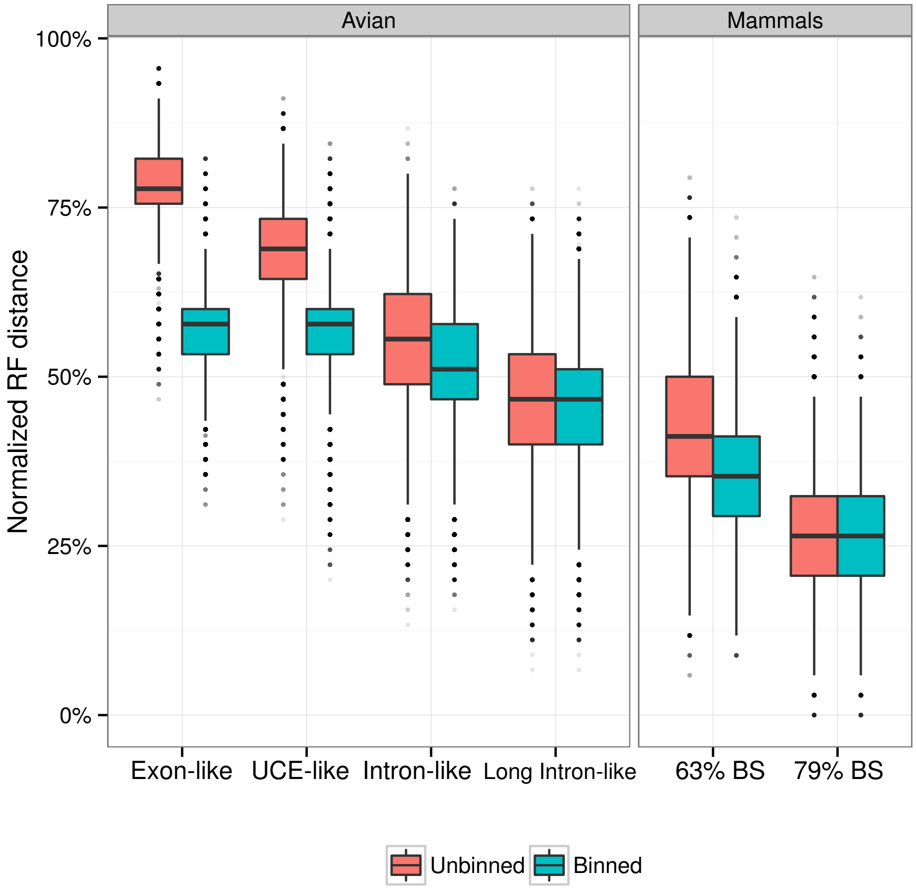


Figure S2: **Gene tree estimation error for various levels of gene tree support for the simulated 1X avian and 1X mammals datasets.** Results are shown for fixed number of genes (1000 for avian and 200 for mammals) and 1X ILS levels. We show the distribution of RF distances between true gene trees and 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.

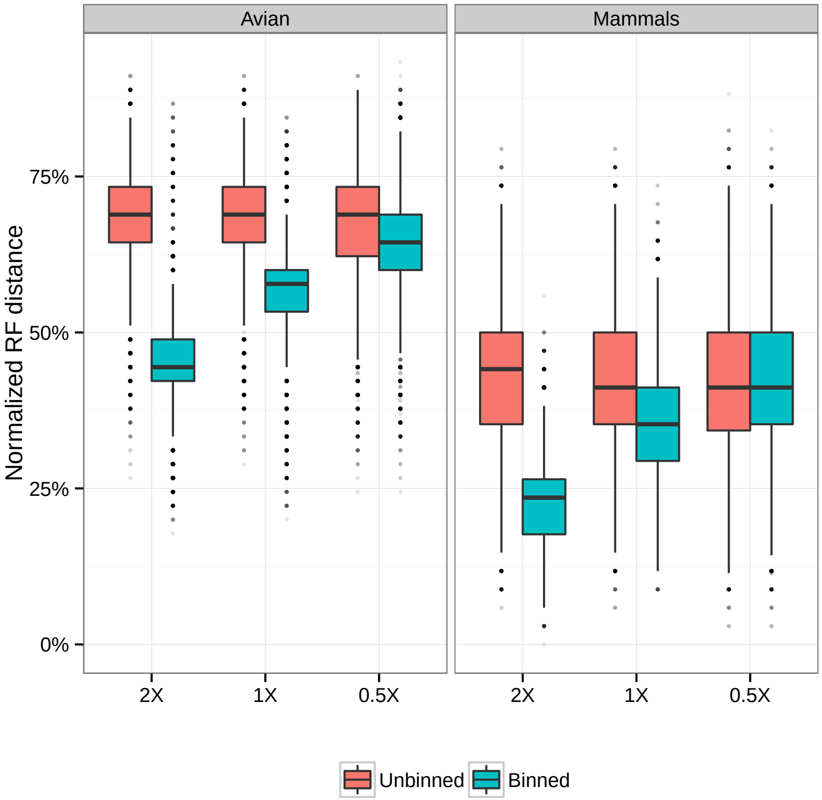


Figure S3: **Gene tree estimation error (computed using RF distances) as a function of ILS level.** We show results for 200 gene trees with 63% BS (moderate phylogenetic signal) for the mammalian simulated datasets, and for 1000 gene trees with UCE-like phylogenetic signal for the avian simulated datasets. 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.

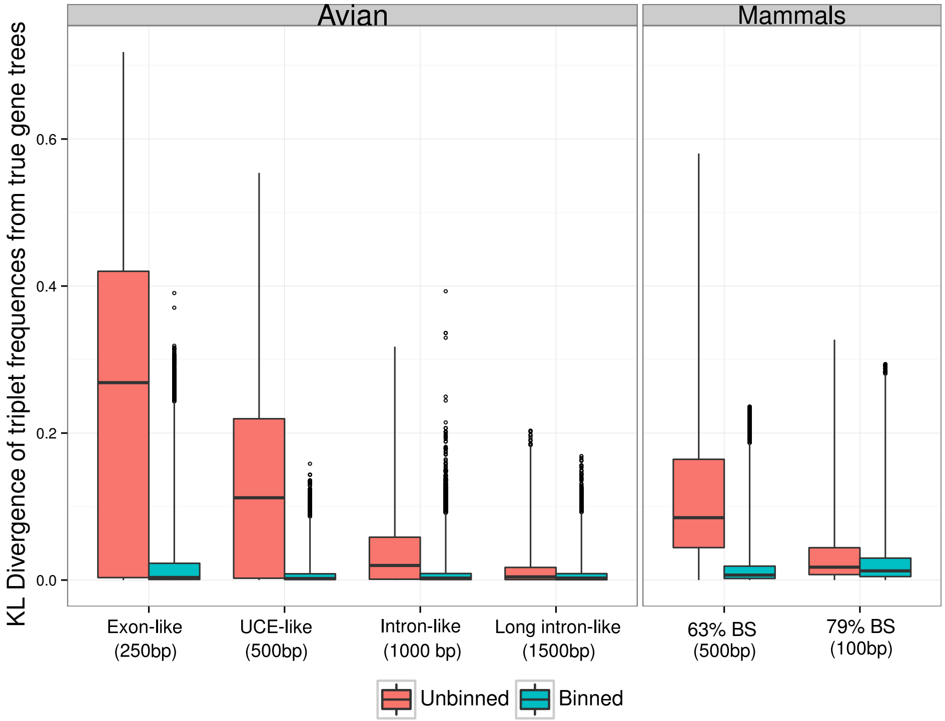


Figure S4: **Divergence of estimated gene trees triplet distributions from true gene tree distributions for simulated avian and mammals datasets with varying levels of gene tree support.** The number of genes is fixed to 1000 for avian and to 200 for mammals, and results are shown only for 1X ILS level. 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 all 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 (*1*).

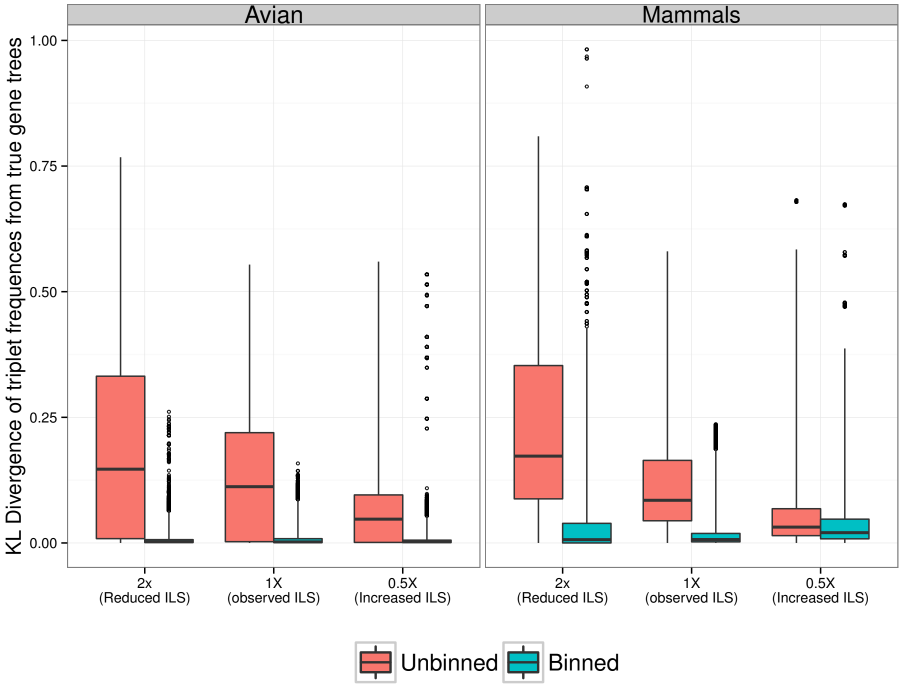


Figure S5: **Divergence of estimated gene tree triplet distributions from true gene tree distributions for simulated avian and mammals datasets with varying levels of ILS.** The number of genes is fixed to 1000 for avian and to 200 for mammals, and gene tree support is fixed to UCE-like for the avian 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 all 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 (*1*).

1.2 Results on simulated datasets

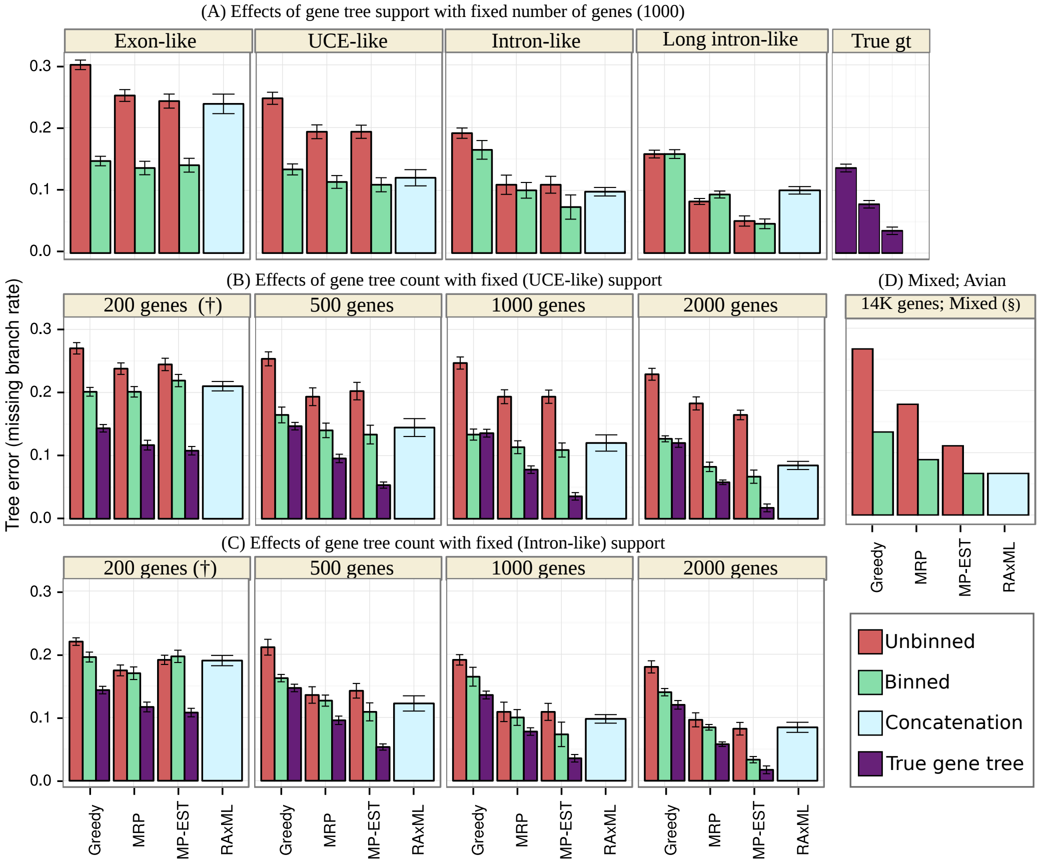


Figure S6: **Simulation results including MRP, Greedy, MP-EST, and concatenation using maximum likelihood on avian datasets.** Bars show average missing branch rates, generally over 10 replicates with some exceptions (†: 20 replicates and §: 1 replicate), and error-bars show standard error. This figure 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 avian dataset with 14350 genes and mixed gene tree support: 8250 exons-like, 2500 intron-like, and 3600 UCE-like.

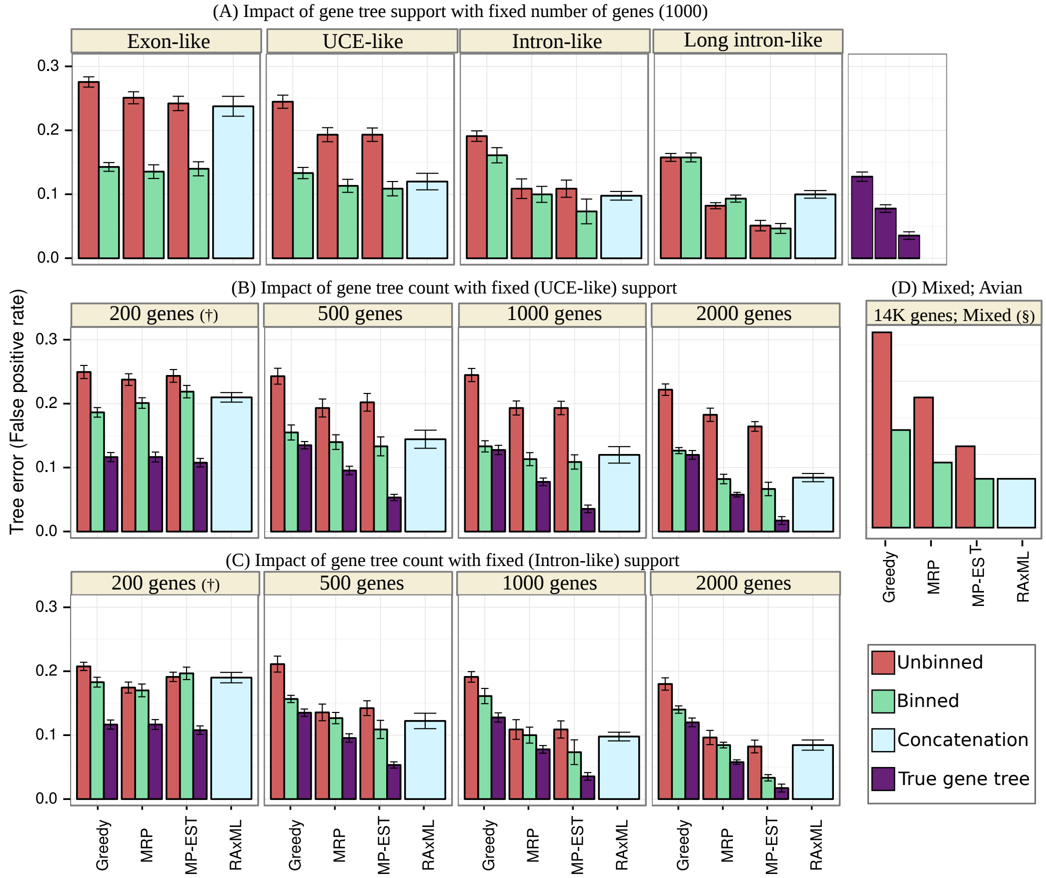


Figure S7: **False positive error rates on simulated avian datasets.** In some rare cases on the simulated avian datasets, the greedy consensus trees produced by the multi-locus bootstrapping procedure were missing one or two edges, and hence had small polytomies. In such cases, the missing branch (false 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. Bars show average false positive rates, generally over 10 replicates with some exceptions (†: 20 replicates, and §: 1 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.

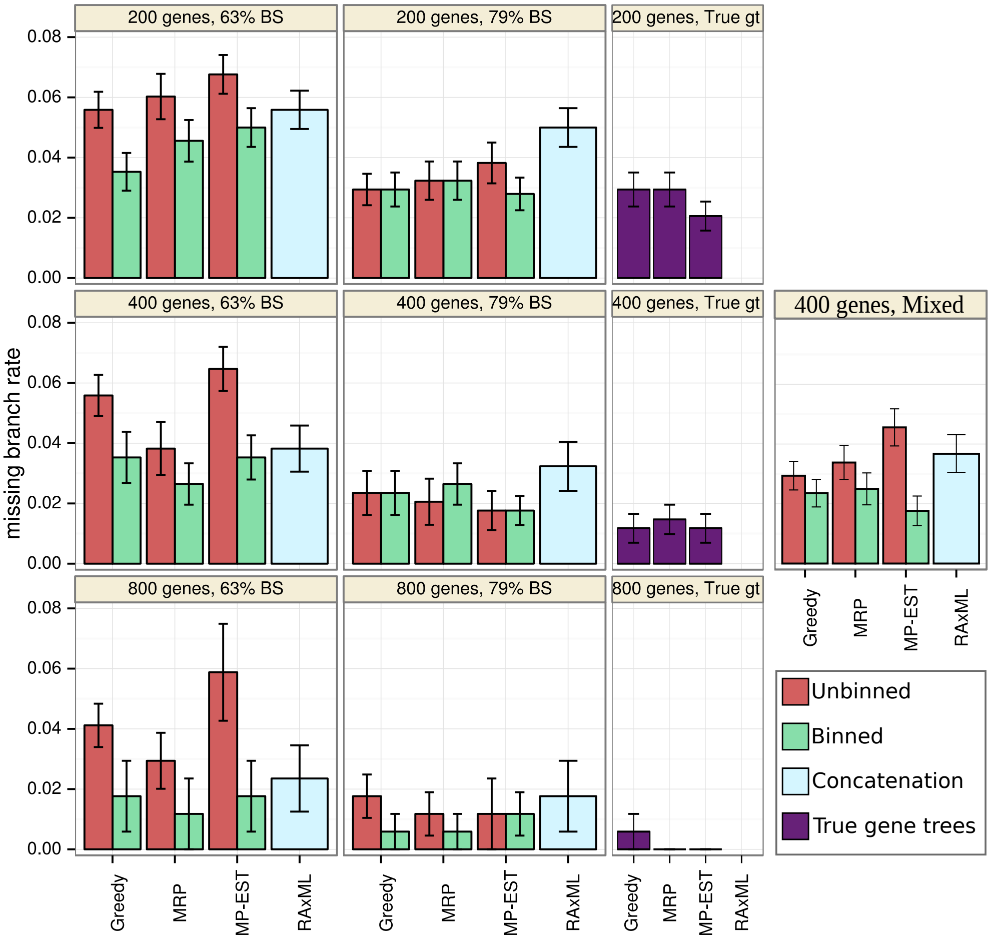


Figure S8: **Simulation results including MRP, Greedy, MP-EST, and concatenation using maximum likelihood on mammalian datasets.** Bars show average missing branch rate over 20 replicates for 200 genes, 10 replicates for 400 genes, and 5 replicates for 800 genes, and error-bars show standard error. This 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).

(1000 genes, 1X ILS)		Exon-like	UCE-like	Intron-like	Long intron-like
Greedy	Unbinned	0.300 (0.024)	0.247 (0.030)	0.191 (0.026)	0.158 (0.010)
Greedy	Binned	0.147 (0.024)	0.133 (0.028)	0.164 (0.047)	0.158 (0.022)
MRP	Unbinned	0.250 (0.030)	0.193 (0.035)	0.109 (0.049)	0.082 (0.015)
MRP	Binned	0.136 (0.034)	0.113 (0.032)	0.100 (0.040)	0.093 (0.018)
MP-EST	Unbinned	0.242 (0.035)	0.193 (0.033)	0.109 (0.043)	0.051 (0.026)
MP-EST	Binned	0.140 (0.035)	0.109 (0.035)	0.073 (0.061)	0.047 (0.024)
RAxML	Concatenation	0.238 (0.049)	0.120 (0.041)	0.098 (0.021)	0.100 (0.019)
(UCE-like, 1X ILS)		200 genes	500 genes	1000 genes	2000 genes
Greedy	Unbinned	0.270 (0.040)	0.253 (0.035)	0.250 (0.030)	0.229 (0.030)
Greedy	Binned	0.201 (0.031)	0.164 (0.039)	0.133 (0.028)	0.127 (0.015)
MRP	Unbinned	0.238 (0.041)	0.193 (0.044)	0.193 (0.035)	0.183 (0.031)
MRP	Binned	0.201 (0.037)	0.140 (0.036)	0.113 (0.032)	0.082 (0.024)
MP-EST	Unbinned	0.244 (0.044)	0.202 (0.044)	0.193 (0.033)	0.164 (0.024)
MP-EST	Binned	0.219 (0.043)	0.133 (0.047)	0.109 (0.035)	0.067 (0.033)
RAxML	Concatenation	0.210 (0.033)	0.144 (0.045)	0.120 (0.041)	0.084 (0.020)
(Intron-like, 1X ILS)		200 genes	500 genes	1000 genes	2000 genes
Greedy	Unbinned	0.220 (0.027)	0.211 (0.040)	0.191 (0.026)	0.180 (0.030)
Greedy	Statistical Binning	0.196 (0.035)	0.162 (0.018)	0.164 (0.047)	0.140 (0.018)
MRP	Unbinned	0.174 (0.038)	0.136 (0.041)	0.109 (0.049)	0.096 (0.033)
MRP	Statistical Binning	0.170 (0.045)	0.127 (0.028)	0.100 (0.040)	0.084 (0.014)
MP-EST	Unbinned	0.191 (0.033)	0.142 (0.037)	0.109 (0.043)	0.082 (0.032)
MP-EST	Statistical Binning	0.197 (0.043)	0.109 (0.046)	0.073 (0.061)	0.033 (0.016)
RAxML	Concatenation	0.190 (0.036)	0.122 (0.038)	0.098 (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)	0.147 (0.019)	0.136 (0.019)	0.120 (0.021)
MRP	True gene tree	0.117 (0.034)	0.096 (0.021)	0.078 (0.019)	0.058 (0.012)
MP-EST	True gene tree	0.110 (0.030)	0.053 (0.016)	0.036 (0.019)	0.017 (0.019)
(1000 genes, UCE-like)		2X	1X	0.5X	
Greedy	Unbinned	0.229 (0.024)	0.247 (0.030)	0.289 (0.021)	
Greedy	Binned	0.073 (0.024)	0.133 (0.028)	0.262 (0.029)	
Greedy	True Gene Trees	0.067 (0.010)	0.136 (0.019)	0.189 (0.046)	
MRP	Unbinned	0.167 (0.019)	0.193 (0.035)	0.222 (0.043)	
MRP	Binned	0.053 (0.019)	0.113 (0.032)	0.164 (0.033)	
MRP	True Gene Trees	0.051 (0.011)	0.078 (0.019)	0.122 (0.024)	
MP-EST	Unbinned	0.162 (0.032)	0.193 (0.033)	0.156 (0.039)	
MP-EST	Binned	0.049 (0.033)	0.109 (0.035)	0.147 (0.028)	
MP-EST	True Gene Trees	0.018 (0.018)	0.036 (0.019)	0.060 (0.030)	
RAxML	Concatenation	0.067 (0.015)	0.120 (0.041)	0.182 (0.042)	

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.

(63% BS; 1X ILS)		200 genes	400 genes	800 genes
Greedy	Unbinned	0.056 (0.027)	0.056 (0.022)	0.041 (0.016)
Greedy	Binned	0.035 (0.028)	0.035 (0.027)	0.018 (0.026)
MRP	Unbinned	0.060 (0.034)	0.038 (0.028)	0.029 (0.021)
MRP	Binned	0.046 (0.031)	0.026 (0.022)	0.012 (0.026)
MP-EST	Unbinned	0.068 (0.029)	0.065 (0.023)	0.059 (0.036)
MP-EST	Binned	0.050 (0.029)	0.035 (0.023)	0.018 (0.026)
RAxML	Concatenation	0.056 (0.028)	0.038 (0.024)	0.024 (0.025)
(79% BS; 1X ILS)		200 genes	400 genes	800 genes
Greedy	Unbinned	0.029 (0.023)	0.024 (0.023)	0.018 (0.016)
Greedy	Binned	0.029 (0.025)	0.024 (0.023)	0.006 (0.013)
MRP	Unbinned	0.032 (0.028)	0.021 (0.024)	0.012 (0.016)
MRP	Binned	0.032 (0.028)	0.026 (0.022)	0.006 (0.013)
MP-EST	Unbinned	0.038 (0.030)	0.018 (0.021)	0.012 (0.026)
MP-EST	Binned	0.028 (0.024)	0.018 (0.015)	0.012 (0.016)
RAxML	Concatenation	0.050 (0.029)	0.032 (0.026)	0.018 (0.026)
(True gene trees)		200 genes	400 genes	800 genes
Greedy	True gene trees	0.029 (0.025)	0.012 (0.015)	0.006 (0.013)
MRP	True gene trees	0.029 (0.025)	0.015 (0.015)	0.000 (0.000)
MP-EST	True gene trees	0.021 (0.022)	0.012 (0.015)	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)
Greedy	True gene trees	0.003 (0.009)	0.029 (0.025)	0.054 (0.027)
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)
MRP	True gene trees	0.003 (0.009)	0.029 (0.025)	0.062 (0.036)
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)
MP-EST	True gene trees	0.003 (0.009)	0.021 (0.022)	0.051 (0.023)
RAxML	Concatenation	0.024 (0.025)	0.056 (0.028)	0.078 (0.037)

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.

Fig.	Model condition	Binned vs. Unbinned	Binned vs. Concat.
Avian simulated datasets			
2A	1X; 1000 genes, varying support	$\mathbf{p < 10^{-5}}$	$\mathbf{p < 10^{-5}}$
2B	1X; UCE-like, varying # genes	$\mathbf{p < 10^{-5}}$	$p = 0.56300$
2C	1X; Intron-like, varying # genes	$\mathbf{p = 0.01580}$	$p = 0.07574$
4A	1k genes; UCE-like, varying ILS	$\mathbf{p < 10^{-5}}$	$\mathbf{p = 0.02504}$
Mammalian simulated datasets			
3A	1X; 63% BS, varying # genes	$\mathbf{p = 0.00112}$	$p = 0.47015$
3A	1X: 79% BS, varying # genes	$p = 0.37182$	$\mathbf{p = 0.00772}$
4B	200 genes; 63% BS, varying ILS	$\mathbf{p = 0.00015}$	$p = 0.24818$

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 (*2*) ($n = 14$), 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 three cases: Fig 2A (1000 avian genes of 1X ILS level, varying gene tree support), 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	**0.00266**	**0.01456**
2B	1X ILS; UCE-like	# genes	**0.01456**	0.76720
2C	1X ILS; Intron-like	# genes	0.13323	0.13323
4A	1k genes; UCE-like	ILS	**0.00036**	0.76720
Mammalian simulated datasets				
3A	1X ILS; 63% BS	# genes	0.76647	0.97895
3A	1X ILS; 79% BS	# genes	0.80057	0.78910
4B	200 genes; 63% BS	ILS	0.32800	0.84538

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 (*2*) ($n = 14$), 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.

1.3 Binning technique

Here we show data related to properties of the binning technique. We first show the impact of threshold t in Figure S9. We then show the properties of the bins created in simulated analyses (Table S6) and biological analyses (Fig. S10). Also, we show results comparing our modification to the Brélaz algorithm to the original Brélaz algorithm (Fig. S11).

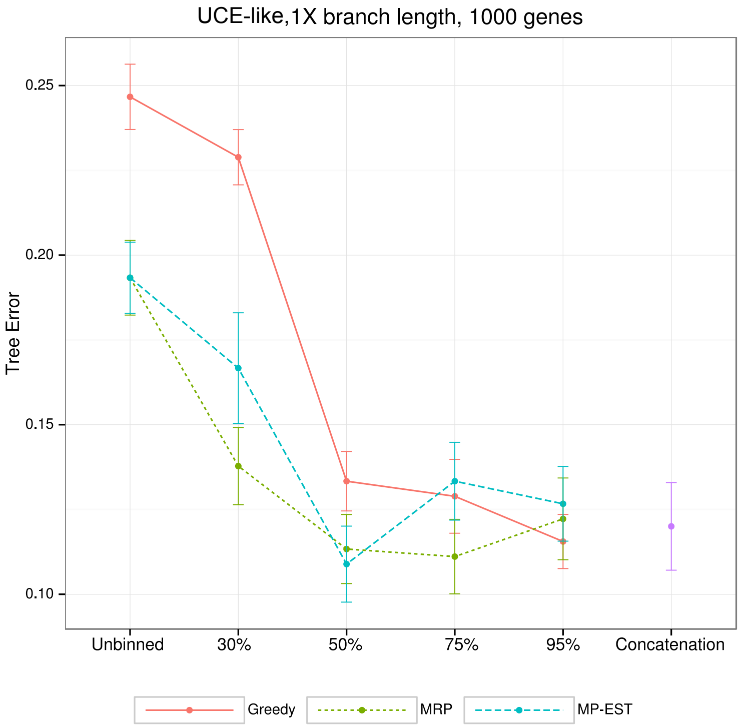


Figure S9: **Effects of support threshold t on the statistical binning.** Results are shown for the simulated avian with 10 replicates, 1000 genes, and UCE-like gene tree support. Dots correspond to average tree error and error bars correspond to standard error. Results are shown for unbinned analyses,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

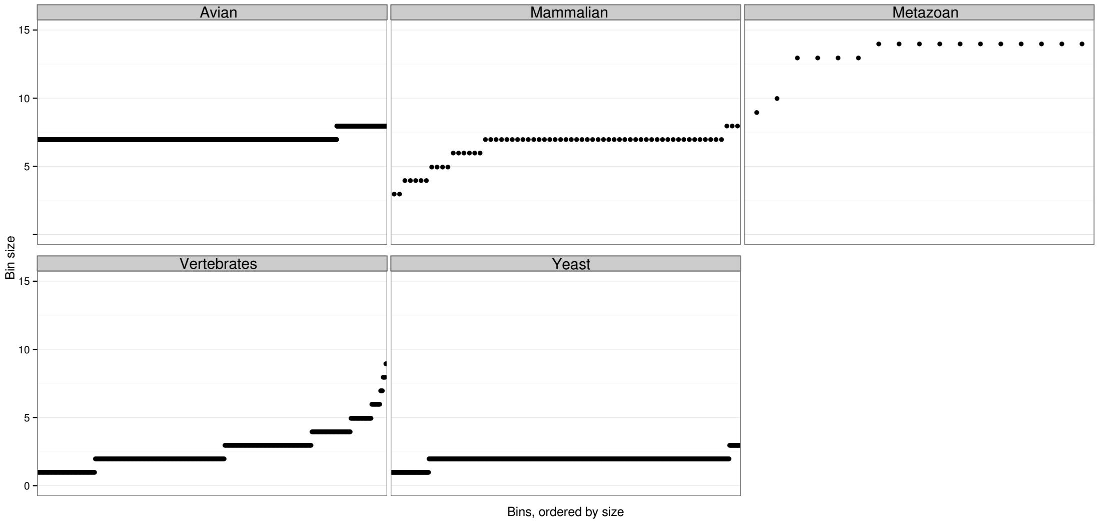


Figure S10: **Bin sizes for the five biological datasets.** On the avian and mammalian datasets almost all bins contain 7 genes (average is 7.1 for avian, and 6.5 for mammalian datasets) on the vertebrate dataset most bins are between 1 and 5 genes and the average is 2.7, on the yeast dataset almost 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.

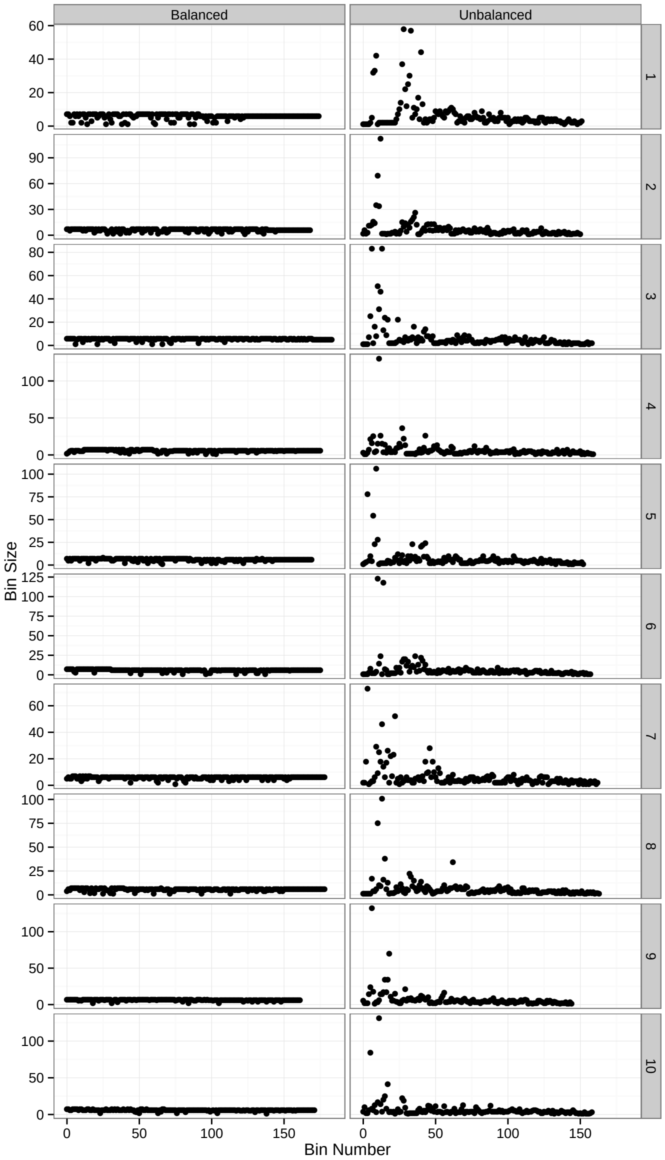


Figure S11: **Bin sizes on the simulated dataset produced by the original Brélaz heuristic ("unbalanced") and our modification ("balanced").** Results are shown for 10 different replicates of the avian simulated UCE-like dataset with 1000 genes and $t = 50\%$. Each dot represents a bin, with vertical axis showing the bins size, and horizontal axis showing the bin index.

Avian Simulated Datasets (Figure 2)				
1000 genes; 1X (Fig. 2A)	Exon-like	UCE-like	Intron-like	Long intron-like
average bin size	15.4	5.7	1.8	1.2
UCE-like; 1X (Fig. 2B)	k=200	k=500	k=1000	k=2000
average bin size	4.2	5.1	5.7	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	1.2	1.3	
Mixed; 1X (Fig. 3B)	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	5.7	2.7	
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.

1.4 Biological results

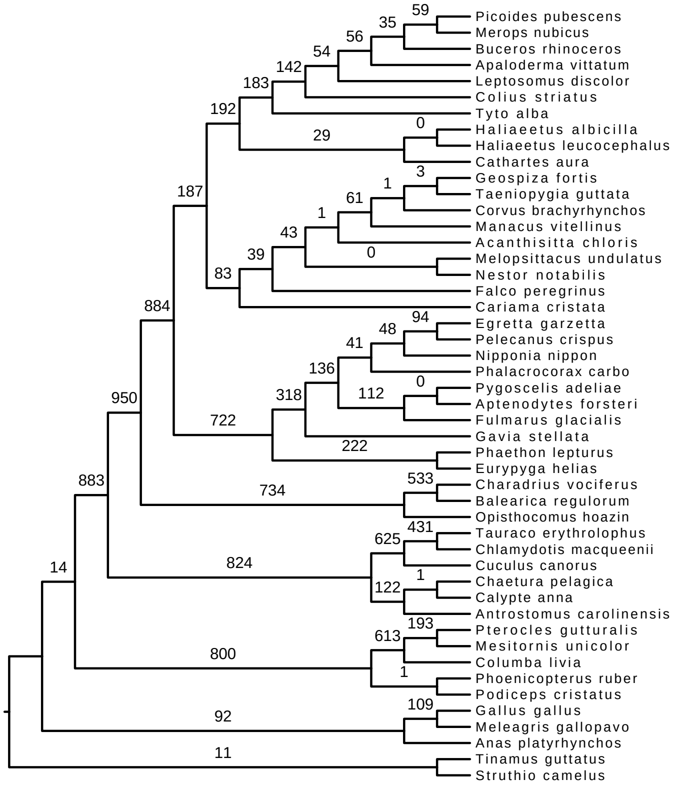


Figure S12: **Evidence for ILS in the avian dataset**. On each branch of the concatenation tree reported in (*3*), we show the number of intron gene trees (out of a total of 2516 loci) that rejected that branch with a BS of at least 75% (these results are also reported by (*3*)). Many of the edges of 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 (*3*), creating a very hard model condition.

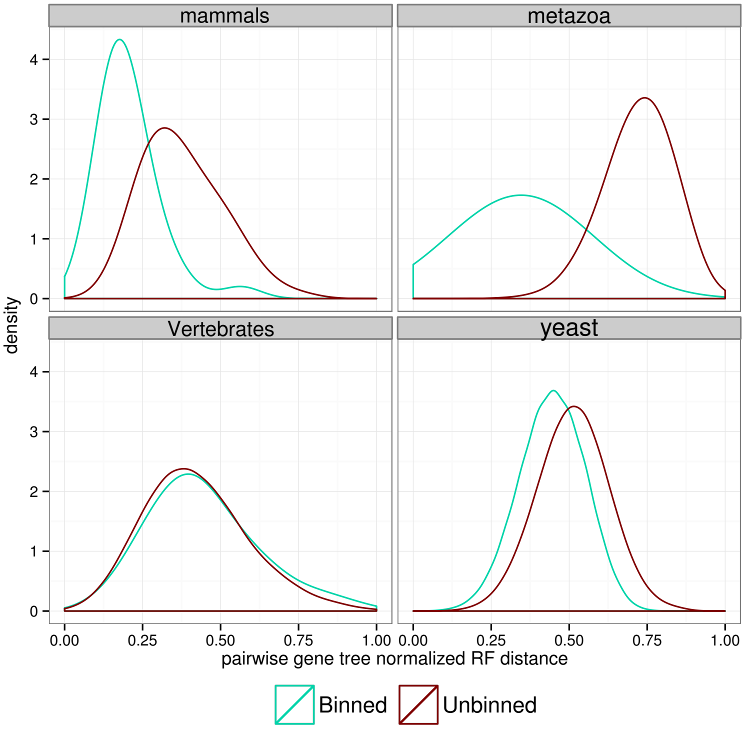


Figure S13: **Gene tree incongruence between pairs of gene trees.** We measure gene tree incongruence using pairwise normalized RF distance between all pairs of gene trees, with and without binning. For biological datasets, the distributions of pairwise gene tree distances are shown as kernel density plots (*4*) using R (*5*). Except for the vertebrate dataset, where binning did not change the amount of incongruence between gene trees, binning reduced incongruence between gene trees.

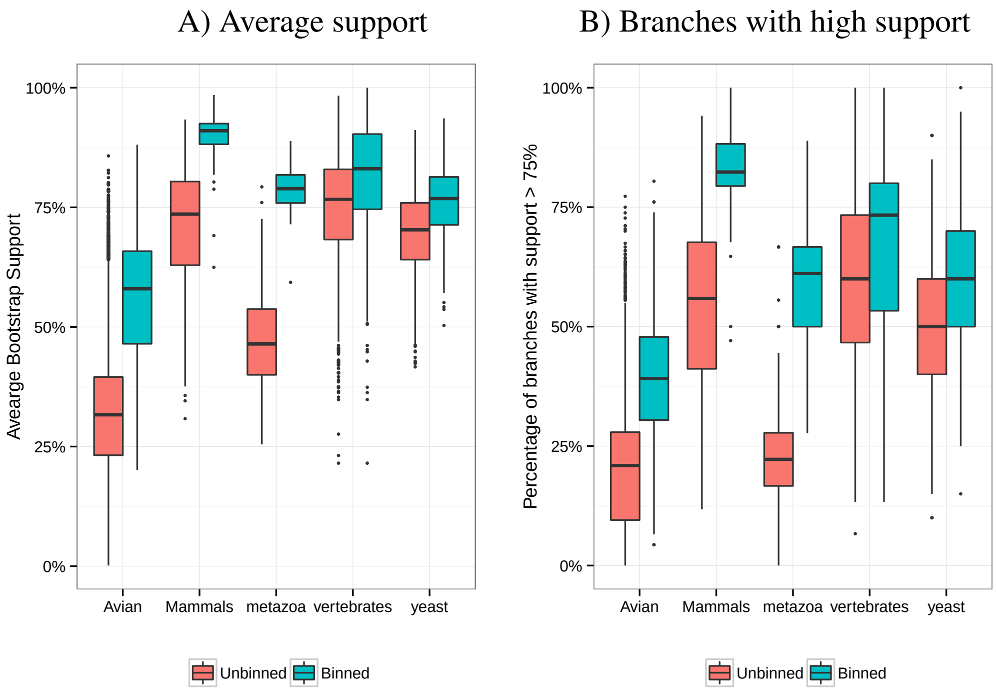


Figure S14: **Gene tree bootstrap support summaries for biological datasets.** A) Boxplots showing distribution of average bootstrap support across all gene trees (solid lines indicate median values). B) Boxplots showing distribution of the percentage of branches in each gene tree that have support above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

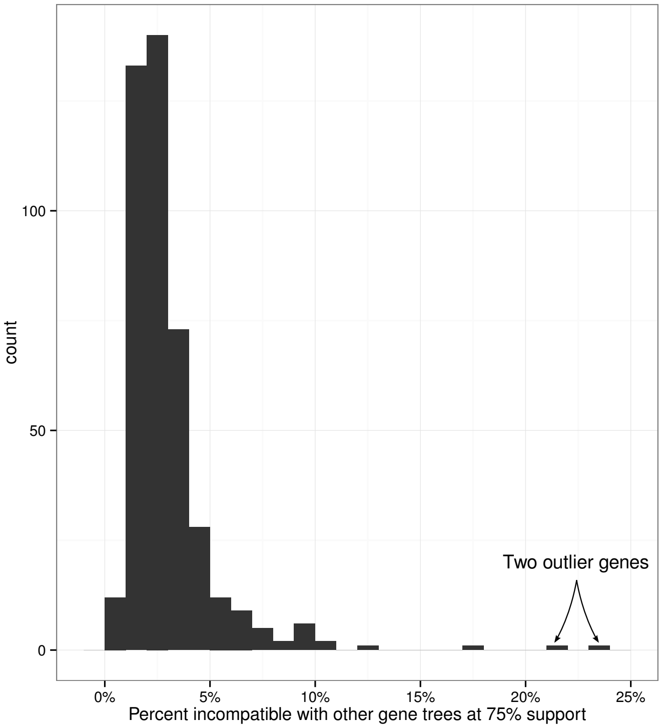


Figure S15: **Outlier mammalian genes.** The histogram shows the distribution of average percentage of branches that each gene had in conflict with other gene trees with at least 75% support. Two genes, with IDs 232 and 209, had on average more than 20% of their edges in conflict with other gene trees 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.

1.4.1 Biological trees

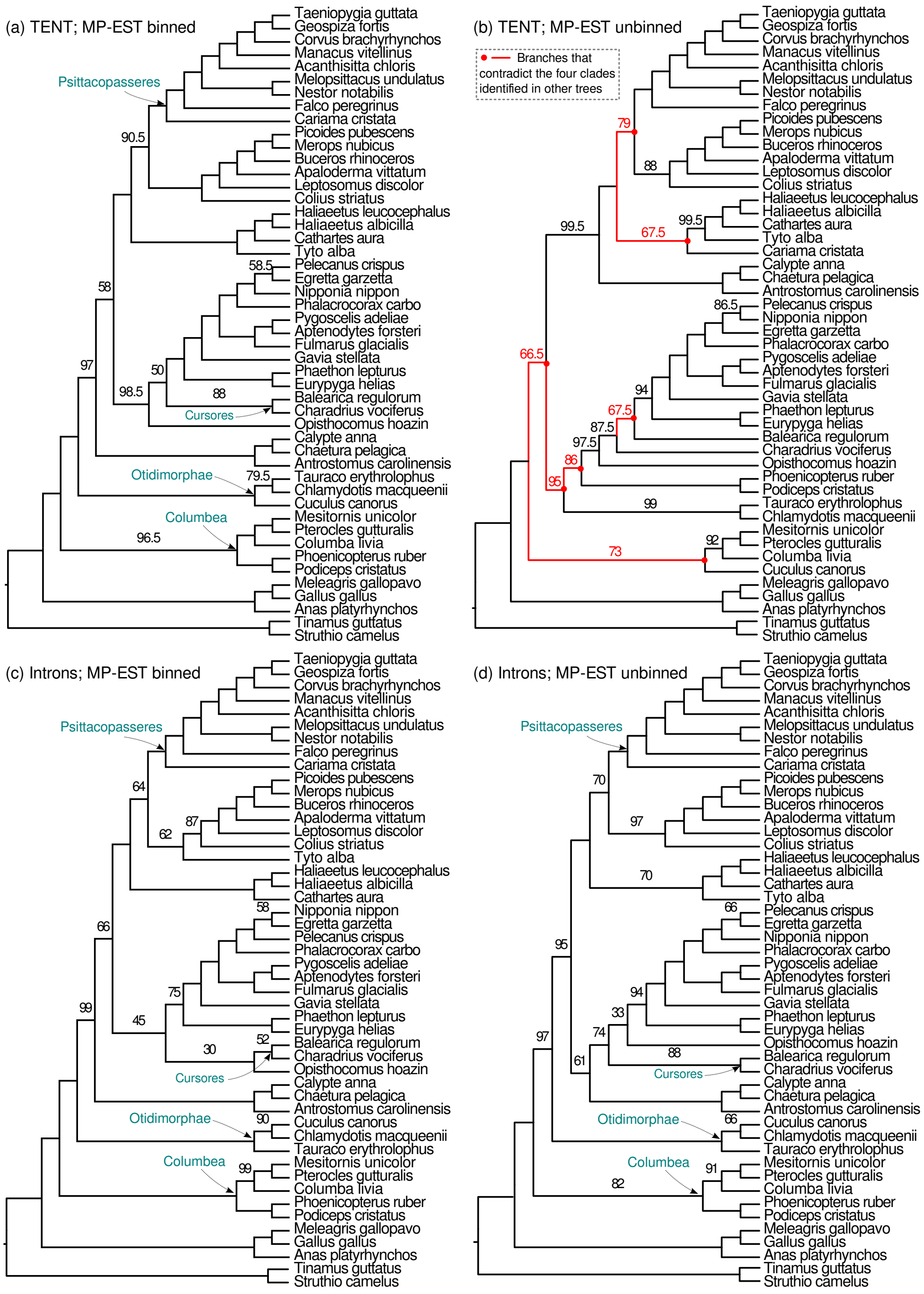


Figure S16: **Binned and unbinned MP-EST trees on the avian dataset.** All trees shown here are reported in (*3*).

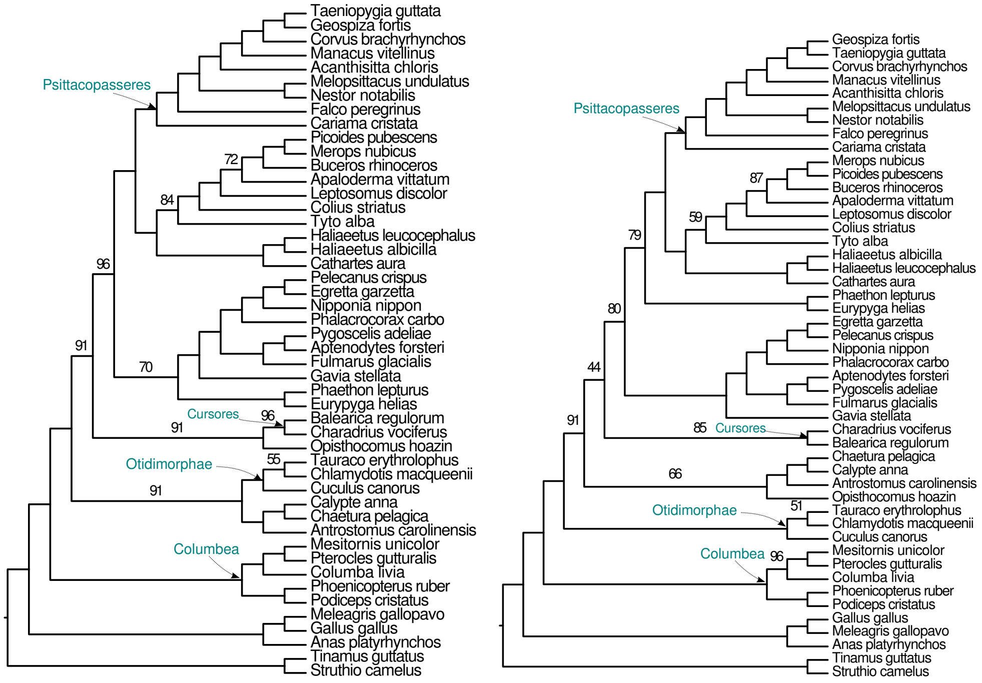
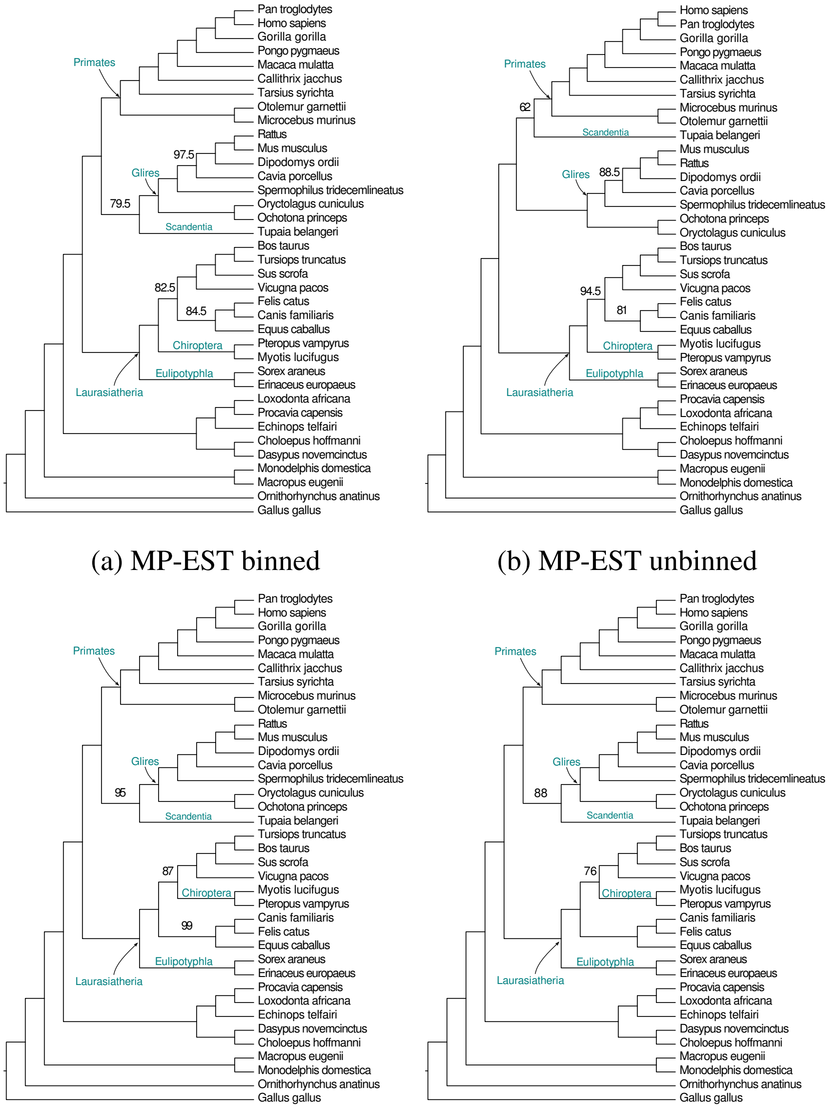


Figure S17: **The concatenation tree reported in (*3*) on the TENT (left) and intron (right) data matrices**. 100% support values are not shown.



(a) MP-EST binned

(b) MP-EST unbinned

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

(d) Concatenation; 424 genes

Figure S18: **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.*, after removing 23 problematic genes; we also show the original concatenated analysis by Song *et al.* on the full set of genes. All edges with no label have 100% support.

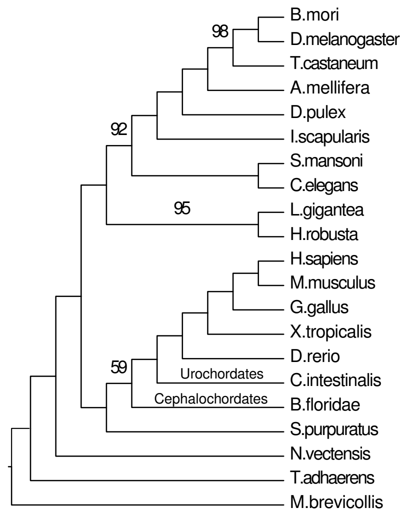


Figure S19: **The concatenation tree reported by Salichos and Rokas (*7*) on the metazoan dataset.** 100% support values are not shown. See the main text for MP-EST trees on the metazoa dataset.

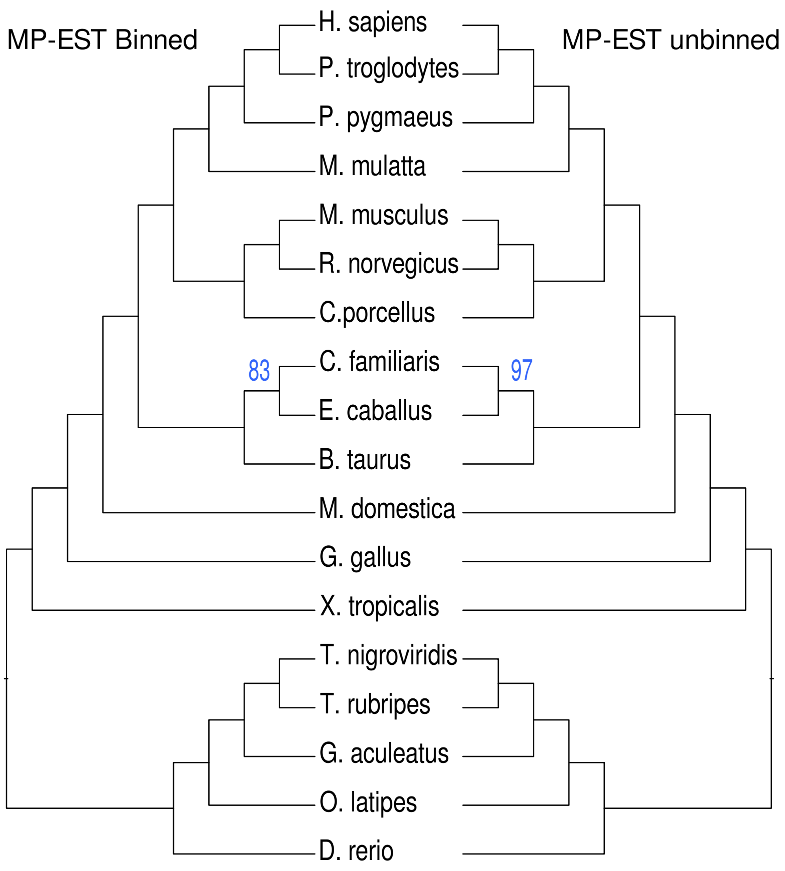
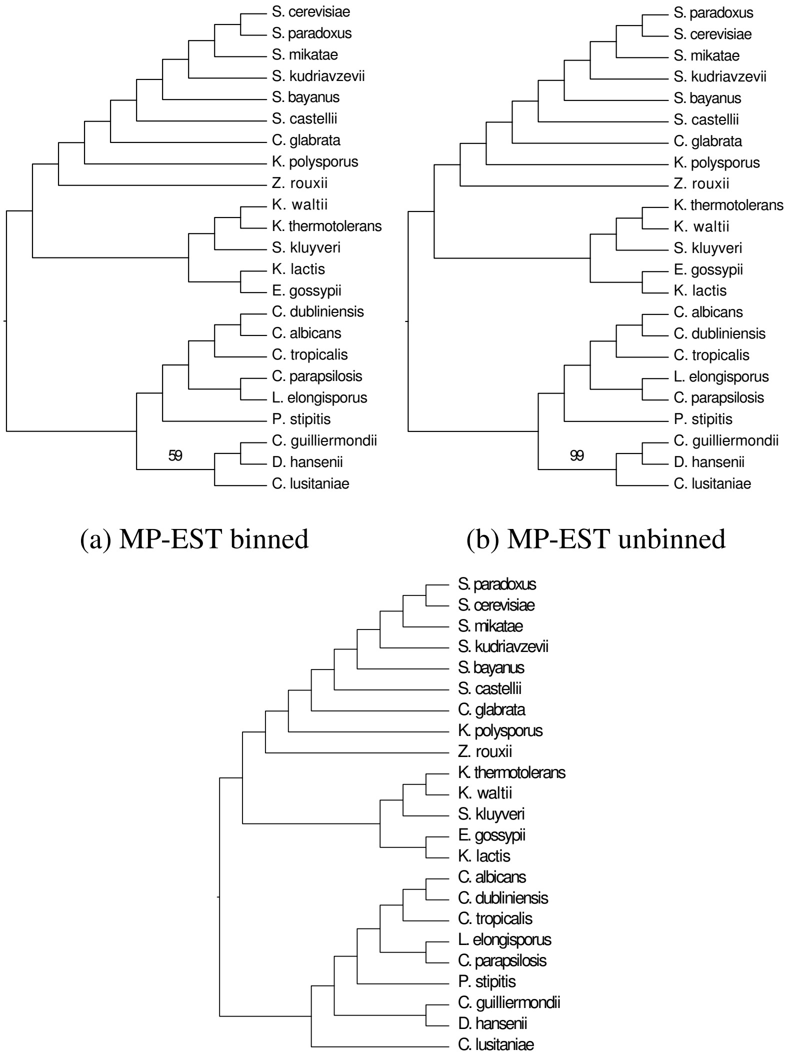


Figure S20: **Binned MP-EST and unbinned MP-EST on the vertebrates dataset.** 100% support values are not shown. Concatenated tree reported by Salichos and Rokas (*7*) is identical to these two trees and has 100% support for all branches.



(a) MP-EST binned

(b) MP-EST unbinned

(c) Concatenation; reported in (*7*)

Figure S21: **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 (*7*). The support on this branch in the unbinned analysis is very high (99%), but much lower (59%) in the binned analysis. The concatenation tree was reported to have 100% support for all edges.

2 Supplementary Discussion

2.1 Biological Datasets

Additional results on the biological datasets are presented in Section 1.4, and discussed here. 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.

2.1.1 Avian dataset analyses

Evidence of ILS: Evidence for ILS in the avian dataset is extensively reported in (*3*). 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 S12, 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 branches are short and 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 (*8*) (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 probable evidence of real gene tree discord.

Species Trees The results on the avian dataset are reported in depth in (*3*). 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 – Psittacopasseres, 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 (*3*), and are discussed in detail in that paper.

Interestingly, both unbinned and binned MP-EST analyses of the intron dataset recover Psittacopasseres, 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 (*3*)). 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.

2.1.2 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 S15). 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 S18 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*).

2.1.3 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 (*9–11*), and is likely correct. However, the assumption before these recent studies was the Cephalochordates/vertebrates relationship (*12*).

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 (*13*), but the majority of recent molecular studies are congruent with the relationship recovered in the concatenation tree, where *N. vectensis* is sister to Bilateria (*14–17*).

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 (*18*). Concatenation (see Fig. S19), 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 (*16, 18, 19*). 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 (*18*). 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 (*20, 21*).

2.1.4 Vertebrate dataset analyses

The binned and unbinned MP-EST trees are topologically identical to each other and to the concatenation tree (Fig. S20); 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 (*22*) for a comprehensive summary.

2.1.5 Yeast dataset analyses

The binned and unbinned MP-EST topologies were identical, and both had 100% support for all but one branch (Fig. S21). Both trees were also identical to the concatenation tree reported in (*7*) 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 (*23*).

2.2 Simulation Studies

We compare the accuracy of species trees estimated using concatenation using maximum likelihood (computed using RAxML (*24*)), and three summary methods (with and without binning). The summary methods we explore are the greedy consensus (also known as the extended majority consensus) (*25*), Matrix Representation with Parsimony (MRP) (*26*), and MP-EST (*27*).

2.2.1 Performance of MRP and Greedy

We begin by discussing results on unbinned summary methods. 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.

However, binning generally improves both Greedy and MRP, similar to the pattern observed for MP-EST. However, there are two cases where binning reduces the accuracy of MRP by a small margin: 400 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 3.3) 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.

2.2.2 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) (*28*) 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.

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. 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.

2.2.3 Concatenation vs. coalescent-based estimation

The relative performance between concatenation and summary methods (whether binned or unbinned) reveals important differences between the avian and mammalian simulation. Hence, we discuss them in turn.

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.

2.3 Statistical Binning Technique

The goal of the statistical binning technique is to produce more accurate gene trees, so that coalescent-based summary methods can have better accuracy. This is achieved through partitioning the set of genes into bins, so that each bin should contain genes with the same evolutionary history. We estimate initial gene trees, and use their topologies to decide if any two genes can be combined. However, as demonstrated, much of the gene tree discordance in the datasets we examined (both biological and simulated) is likely due to estimation error, for example caused by insufficient sequence length. Thus, we have chosen to consider two genes incombinable only if they show discordance at a chosen support threshold, and we treat weakly supported discordances as potentially due to estimation error. Nevertheless, some of these weakly supported discordances could be real, and hence the set of genes put in the same bin can in fact have real incongruence. But this possibility need not necessarily discourage us from forming bins. The key observation is to realize that the inputs to the summary methods are estimated gene trees, and hence will always have estimation error, whether binning is performed or not. If the accuracy gain resulting from increased phylogenetic signal is greater than the error introduced by placing genes with different phylogenies in the same bin, then forming bins will be beneficial, because the overall error is reduced.

As the results have shown, binning improves the accuracy of estimated gene trees and triplet gene tree distributions, and thus leads to improvement in species tree topology and branch length estimation. The cases where this impact is strongest is for the genes with low to moderate phylogenetic signal, where binning can reduce error substantially. However, when genes have very strong phylogenetic signal (the better mammalian genes in the simulation or the long intron-like avian genes), then the impact of binning is neutral or only minimal.

3 Methods

3.1 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 (*29, 30*). 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 (*31*).

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 then 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, any 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 (*32*), and so cannot currently be solved exactly for large graphs. However, there are heuristics for finding good vertex colorings (*33,34*). 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 (see Section 2): 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 (*33*), 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 developed by Shalin Shah (*35*). 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 S11.

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.).

3.2 Simulation study protocol

We simulated two sets of datasets, one based on the avian dataset with 48 species and 14,446 loci studied in (*3*), and one based on the mammals dataset with 37 species and 447 loci studied in (*6*).

The Jarvis *et al.* 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. Alignments were computed on each of these regions using SATé-2 (*36*), and RAxML under GTRGAMMA was used to compute bootstrapped gene trees. 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 Song *et al.* dataset had 37 mammalian 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.

We created model conditions that resembled the biological datasets from Jarvis *et al.* and Song *et al.* 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. S18) 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 (*3*), 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 (*37*) and based on the multi-species coalescent model.

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 (*38*) based on parameters estimated by bppml (*38*) 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 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 (10 replicates), 1000 genes (10 replicates), and 2000 genes (10 replicates) for the avian, and 200 genes (20 replicates), 400 genes (10 replicates), and 800 genes (5 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.

3.3 Multi-locus bootstrapping

We use the site-only multi-locus bootstrapping (*39*) 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.

3.4 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.

3.5 Concatenated analyses

The concatenated 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. Concatenated 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.

3.6 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) (*40*) 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 (*28*) 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.

3.7 Commands and version numbers

3.7.1 Estimating ML gene trees

We used RAxML version 7.3.5 (*24*) 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]
```

3.7.2 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.

3.7.3 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) (*41*) 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;
```

3.7.4 Greedy

We use Dendropy version 3.12.0 (*37*) to compute the greedy consensus tree.

3.7.5 Binning

A set of custom shell and Python (*42*) scripts are used to build the binning pipeline. The entire pipeline will be made publicly available (in open source form) at the time of publication. Each step is described below.

Finding compatibility between pairs of gene trees was performed using Phylonet version 2.4.0 (*43*), with custom modifications to the code to 1) handle missing taxa correctly and 2) compare one gene against multiple genes in one run. The following command was used:

```
 java -jar phylonet.jar compat tree1 tree2 b
```

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 (*35*). 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.

3.8 Computational resources and issues

3.8.1 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 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 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.

3.8.2 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 10-20 replicates. 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.

3.8.3 Summary of running time considerations

Using summary methods (whether binned or unbinned) that depend on MLBS (multi-locus bootstrapping) is computationally intensive when there are many genes, because running RAxML with bootstrapping is computationally intensive, and it must be applied to every gene. Alternative and faster techniques for maximum likelihood tree estimation exist, such as FastTree-2 (*44, 45*), but in this study we used RAxML, which is better known. Reductions in running time could also potentially be achieved through alternative techniques to bootstrapping, surveyed in (*46*).

MP-EST is itself computationally intensive because it attempts to solve maximum likelihood under the coalescent model, and so is impacted by the number of taxa. If a fast summary method is used, such as Greedy or MRP, then almost 100% of the running time is spent computing gene trees with bootstrapping. If MP-EST is used instead, then the running time can easily double, but the running time depends on the number of taxa. Thus, on the 2000-gene avian datasets, the running time for using unbinned MP-EST roughly doubled compared to just computing the maximum likelihood gene trees with bootstrapping.

Binning increases the time because in addition to computing gene trees (with bootstrapping), it also computes supergene trees, and this can take up to as much time as gene tree computation. Hence, a binned analysis using Greedy or MRP can be almost twice as long as an unbinned analysis with the fast summary method; however, a binned analysis using MP-EST uses 50% more time than an unbinned analysis using MP-EST because MP-EST is used only once, and is itself very expensive.

References

1. M. Hubert, E. Vandervieren, *Computational Statistics & Data Analysis* **52**, 5186 (2008).

2. Y. Benjamini, Y. Hochberg, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 289–300 (1995).

3. E. Jarvis, *et al.* (2013). Avian phylogenomics paper, in preparation for submission to Science.

4. B. Silverman, *Monographs on Statistics and Applied Probability*, no. 1951 (Chapman & Hall/CRC, 1986), p. 176.

5. R. R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing (2011). ISBN 3-900051-07-0.

6. S. Song, L. Liu, S. V. Edwards, S. Wu, *Proceedings of the National Academy of Sciences of the United States of America* **109**, 14942 (2012).

7. L. Salichos, A. Rokas, *Nature* **497**, 327 (2013).

8. N. Rosenberg, *Molecular Biology and Evolution* **30**, 2709 (2013).

9. F. Delsuc, H. Brinkmann, D. Chourrout, H. Philippe, *Nature* **439**, 965 (2006).

10. S. J. Bourlat, *et al.*, *Nature* **444**, 85 (2006).

11. T. R. Singh, *et al.*, *BMC Genomics* **10**, 534 (2009).

12. C. Nielsen, *Animal evolution: interrelationships of the living phyla* (Oxford University Press, 2012).

13. B. Schierwater, *et al.*, *PLoS Biology* **7**, e1000020 (2009).

14. A. Hejnol, *et al.*, *Proceedings of the Royal Society B: Biological Sciences* **276**, 4261 (2009).

15. E. A. Sperling, K. J. Peterson, D. Pisani, *Molecular Biology and Evolution* **26**, 2261 (2009).

16. H. Philippe, *et al.*, *Current Biology* **19**, 706 (2009).

17. J. F. Ryan, *et al.*, *Science* **342**, 1242592+ (2013).

18. G. D. Edgecombe, *et al.*, *Organisms Diversity & Evolution* **11**, 151 (2011).

19. N. Lartillot, H. Philippe, *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences* **363**, 1463 (2008).

20. K. Ishiwata, G. Sasaki, J. Ogawa, T. Miyata, Z.-H. Su, *Molecular Phylogenetics and Evolution* **58**, 169 (2011).

21. S. J. Longhorn, H. W. Pohl, A. P. Vogler, *Molecular Phylogenetics and Evolution* **55**, 846 (2010).

22. J.-Y. Hu, Y.-P. Zhang, L. Yu, *Dongwuxue Yanjiu* **33**, E65 (2012).

23. B. Dujon, *Nature Reviews, Genetics* **11**, 512 (2010).

24. A. Stamatakis, *Bioinformatics* **22**, 2688 (2006).

25. J. Degnan, M. DeGiorgio, D. Bryant, N. Rosenberg, *Systematic Biology* **58**, 35 (2009).

26. M. A. Ragan, *Molecular Phylogenetics and Evolution* **1**, 53 (1992).

27. L. Liu, L. Yu, S. V. Edwards, *BMC Evolutionary Biology* **10**, 302 (2010).

28. R. A. Leibler, S. Kullback, *The Annals of Mathematical Statistics* **22**, 79 (1951).

29. E. Wilson, *Systematic Zoology* **14**, 214 (1965).

30. W. J. LeQuesne, *Systematic Zoology* **18**, 201 (1969).

31. T. Warnow, *Journal of Algorithms* **16**, 388 (1994).

32. E. Malaguti, P. Toth, *International Transactions in Operational Research* **17**, 1 (2010).

33. D. Brélaz, *Communications of the ACM* **22**, 251 (1979).

34. J. C. Culberson, Iterated greedy graph coloring and the difficulty landscape, *Tech. rep.*, University of Alberta (1992).

35. S. Shah, Webpage. `http://shah.freeshell.org/graphcoloring/`.

36. K. Liu, *et al.*, *Systematic Biology* **61**, 90 (2012).

37. J. Sukumaran, M. Holder, *Bioinformatics* **26**, 1569 (2010).

38. J. Dutheil, B. Boussau, *BMC Evolutionary Biology* **8**, 255 (2008).

39. T.-K. Seo, *Molecular Biology and Evolution* **25**, 960 (2008).

40. D. Robinson, L. Foulds, *Mathematical Biosciences* **53**, 131 (1981).

41. D. Swafford, *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4.*, Sunderland, Massachussets (2002).

42. G. van Rossum, Python tutorial, *Tech. Rep. CS-49526*, Centrum voor Wiskunde en Informatica (CWI), Amsterdam (1995).

43. C. Than, D. Ruths, L. Nakhleh, *BMC Bioinformatics* **9**, 322 (2008).

44. M. Price, P. Dehal, A. Arkin, *PLoS ONE* **3**, e9490 (2010).

45. K. Liu, C. R. Linder, T. Warnow, *PLoS ONE* **6**, e27731 (2012). Doi:10.1371/journal.pone.0027731.

46. M. Anisimova, M. Gil, J. Dufayard, C. Dessimoz, O. Gascuel, *Systematic Biology* **60**, 685 (2011).