Weighted Statistical Binning: enabling statistically consistent genome-scale Phylogenetic Analyses

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

1 Department of Computer Science, University of Texas at Austin, Austin, Texas, USA

2 Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA

∗ **E-mail: warnow@illinois.edu**

Abstract

Because biological processes can make different loci have different evolutionary histories, species tree estimation requires multiple loci from across the genome. While many processes can result in discord between gene trees and species trees, incomplete lineage sorting (ILS), modeled by the multi-species coalescent, is considered to be a dominant cause for gene tree heterogeneity. Coalescent-based methods have been developed to estimate species trees, many of which operate by combining estimated gene trees, and so are called "summary methods". Because summary methods are generally fast (and much faster than more complicated coalescent-based methods that co-estimate gene trees and species trees), they have become very popular techniques for estimating species trees from multiple loci. However, recent studies have established that summary methods can have reduced accuracy in the presence of gene tree estimation error, and also that many biological datasets have substantial gene tree estimation error, so that summary methods may not be highly accurate on biologically realistic conditions. Mirarab et al. (Science 2014) presented the "statistical binning" technique to improve gene tree estimation in multi-locus analyses, and showed that it improved the accuracy of MP-EST, one of the most popular coalescent-based summary methods. Statistical binning, which uses a simple statistical test for "combinability" and then uses the larger sets of genes to re-calculate gene trees, has good empirical performance, but using statistical binning within a phylogenomics pipeline does not have the desirable property of being *statistically consistent*. We show that weighting the re-calculated gene trees by the bin sizes makes statistical binning statistically consistent under the multispecies coalescent, and maintains the good empirical performance. Thus, "weighted statistical binning" enables highly accurate genome-scale species tree estimation, and is also statistical consistent under the multi-species coalescent model.

Introduction

The estimation of phylogenetic trees, whether of individual loci (so called "gene trees") or at the genome-level (species trees), is a basic step in many biological analyses, with many applications in biological data analysis [1]. However, estimating gene trees and species trees with high accuracy is difficult for many reasons, including computational issues (nearly all problems are NP-hard) and dataset issues. For example, while highly accurate gene trees can be computed for some loci, when a locus has limited *phylogenetic signal* (e.g., its sequences are too short, or it evolves too slowly), its gene tree may only be estimated with partial accuracy. Species tree estimation is also difficult, because different loci can have different phylogenetic trees, a phenomenon that occurs due to several different biological processes. In particular, many groups of species evolve with rapid speciation events, a process that is likely to produce conflict between gene trees and species trees due to *incomplete lineage sorting* (ILS) [2–5]. Furthermore, when ILS occurs, standard methods for estimating species trees, such as concatenation (which combines sequence alignments from different loci into a single "super-alignment", and then computes a tree on the super-alignment) and consensus methods, can be statistically inconsistent [6, 7], and produce highly supported but incorrect trees [8]; also, gene tree heterogeneity and poor phylogenetic signal generally makes phylogeny estimation difficult [9]. Because standard methods can be positively misleading in the presence of gene tree heterogeneity due to ILS, statistical methods (e.g., [10–14]) have been developed to estimate the species tree in the presence of gene tree heterogeneity. Many of these methods operate by combining estimated gene trees (and so are called "summary methods"), and can be fast enough to analyze datasets with hundreds to thousands of genes. One of the important properties for a summary method is *statistical consistency under the multi-species coalescent model*, which means that the probability that the method will return the true species tree will converge to 1, as the number of gene trees in the input increases (under the assumption that discord between the species tree and the gene trees is due only to incomplete lineage sorting). Of the many statistically consistent methods, MP-EST [15], a maximum pseudo-likelihood method, is one of the most popular. However, empirical and simulation studies have shown that summary methods (such as MP-EST) are impacted by gene tree estimation error, and can produce less accurate species trees than concatenation when gene tree estimation error is high enough [16–18] (and see also discussion in [19]). In a genome-scale analysis, it is unlikely that all the loci will have substantial phylogenetic signal, and so this vulnerability to gene tree estimation error potentially means that coalescent-based summary methods may not be highly accurate techniques for estimating species trees from genome-scale data. This is particularly problematic when the sequences for each locus are kept short enough to avoid recombination [17, 19], which also violates the assumptions of the multi-species coalescent model, since short sequences will tend to have insufficient phylogenetic signal to provide full resolution of the gene trees (however, see [20]).

This vulnerability presented a significant analytical challenge to the avian phylogenomics project, which estimated a species tree for modern birds using whole genomes for 48 avian species [21] (see also [19] about how gene tree estimation error may have caused problems for coalescent analyses of a mammalian biological dataset). The project computed a maximum likelihood concatenation analysis and compared the concatenation tree to the estimated gene trees for 14,446 different loci. These gene trees were all topologically different from each other, and also from the concatenation analysis, indicating tremendous gene tree heterogeneity. Furthermore, the concatenation analysis indicated multiple speciation events in rapid succession, which increases the probability that ILS will be present [5]. Together, these observations suggest that the avian species tree has substantial ILS, and so a coalescent-based analysis was needed. Jarvis *et al.* [21] used MP-EST with multi-locus bootstrapping (MLBS [22]) to estimate a species tree, but the tree they obtained was very different from the concatenation analysis, raising questions about which analysis was more reliable. Since the MP-EST tree also violated several subgroups established in various analyses (including Australaves, which has been supported in multiple studies [23–26]), they suspected that the MP-EST tree was less likely to be correct. In addition, the 14,446 gene trees showed very low bootstrap support (average support $\leq 40\%$), and so the most likely explanation for the difference between the coalescent-based species tree produced by MP-EST and the concatenation analysis was gene tree estimation error.

Mirarab et al. [18] developed a technique they called "Statistical Binning" to improve species tree estimation using coalescent-based species tree estimation methods. Statistical binning partitions the genes into sets, based on a statistical test for "combinability", concatenates the gene sequence alignments within each set into a "supergene alignment", and then estimates a "supergene tree" on the supergene alignment. The newly estimated supergene trees can be used by a coalescent-based summary method to compute a species tree on the dataset. As shown in [18], statistical binning improves the estimation of gene trees and gene tree distributions, and this results in improved species tree topologies and branch lengths when species trees are computed using MP-EST on MLBS gene trees. Furthermore, MP-EST, used with statistical binning, was almost always at least as accurate as concatenation (more accurate than concatenation when the ILS level is high, and only less accurate than concatenation for very low levels of ILS). Finally, statistical binning followed by MP-EST was used to compute a species tree on the avian phylogenomic dataset, and this "MP-EST*" tree was nearly identical to the concatenation analysis they obtained; the MP-EST* and concatenation trees are presented in [21] as the two major hypothesis for the avian phylogeny.

Weighted Statistical Binning

Statistical binning performed well on the simulated and biological data studied in [18], but the use of statistical binning has some drawbacks that can be important, as we will show. Specifically, the technique used to bin the genes into disjoint subsets could, under some conditions, produce a set of supergene trees that no longer accurately estimates the distribution of true gene trees, and when this happens the species tree estimation pipeline will not be statistical consistent under the multi-species coalescent model (Theorem 1). However, we will also show a simple modification to the statistical binning technique that makes it statistically consistent under the multi-species coalescent model, when followed by a statistically consistent coalescent-based summary method, such as MP-EST (Theorem 2). Furthermore, we will present the results on both simulated and biological data that demonstrate that this new technique, which we call "weighted statistical binning", produces highly accurate species trees and branch lengths, just like (unweighted) statistical binning. Hence, our study suggests that weighted statistical binning has the same (or better) empirical advantages of unweighted statistical binning, and also proves that weighted statistical binning is statistically consistent under the multi-species coalescent model.

The statistical binning technique presented in [18] operates as follows. The input is a multiple sequence alignment and phylogenetic tree (with bootstrap support on its branches) on each of n given genes, and a user-specified "threshold support" value $t < 1$. The role of the threshold t is to specify which branches in the gene trees are considered reliable, and which ones have support that is so low that the branches may be due to estimation error. Therefore, if the trees on two genes differ only in terms of their low support edges, the differences are considered potentially consistent with estimation error, and the two genes are considered "combinable" or "compatible".

Statistical binning uses a simple statistical heuristic to determine which pairs of gene trees cannot be put into the same bin; this test is based on the bootstrap support on the branches, and will prevent two genes from being in the same bin if their gene trees have conflicting branches, each with bootstrap support of at least t. This is the combinability test, so that two genes are not considered combinable if they have highly supported conflicting branches, and otherwise are considered combinable. (Equivalently, two genes are combinable if their gene trees, after collapsing all the low support branches, share a common refinement.) Finally, because pairwise compatibility ensures setwise compatibility [27], if a set of gene trees *can* be all put in the same bin, then there is a tree that combines all the highly supported branches in any of the trees in the set.

The first step in statistical binning creates a graph based on the input, and uses a graph-theoretic optimization to bin the genes into subsets. Each gene is represented by a single vertex in the graph, and an edge is placed between two genes if their gene trees are not combinable, based on the statistical test described above. Determining if two genes are combinable can be computed in linear time [28], and so this graph, which we call the incompatibility graph, can be computed in time linear in the number of taxa and quadratic in the number of genes. The vertices of the graph are then partitioned into subsets, so that there are no edges between any two vertices in the same subset (i.e., all the genes in the subset are combinable). To partition the vertices, the vertices are assigned colors so that no two vertices with the same color are adjacent; then, all vertices with the same color are placed in the same bin. After the bins are computed, the sequence alignments for the genes in a given bin are concatenated, and a tree is computed on the concatenated alignment. Since longer sequences tend to produce more accurate gene trees [29], having the bins be as large as possible is desirable; this is accomplished by seeking a coloring with as few colors as possible (i.e., a minimum vertex coloring), which is an NP-hard problem [30]. However, summary methods, such as MP-EST, use the distribution of the gene trees to estimate the species tree; hence, balanced bins (of the same size) are desirable so that the distribution of supergene trees is close to the input gene tree distribution. Therefore, the objective is a coloring of the vertices, using a small number of colors, so that every color class contains about the same number of colors. To achieve such a coloring, [18] modified the Brélaz heuristic [31] for minimum vertex coloring, so that during the greedy coloring, a node is added to the smallest bin for which it has no conflicts.

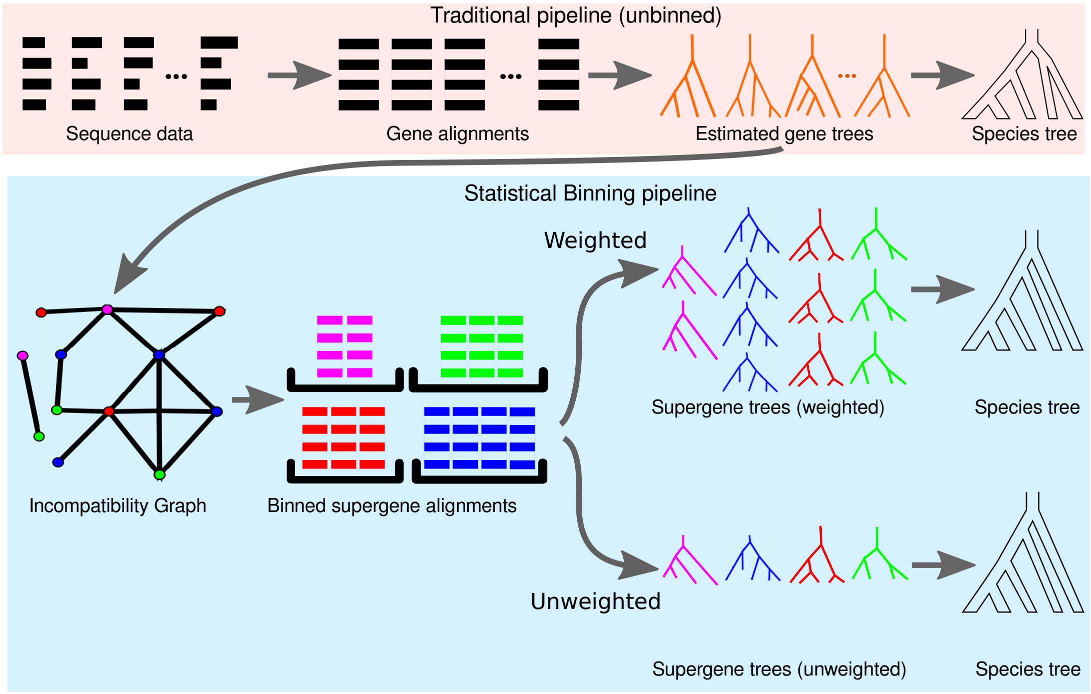

Traditional pipeline (unbinned)
Sequence data
Gene alignments
Estimated gene trees
Species tree
Statistical Binning pipeline
Weighted
Incompatibility Graph
Binned supergene alignments
Supergene trees (weighted)
Species tree
Unweighted
Supergene trees (unweighted)
Species tree

Figure 1. Pipeline for unbinned analyses, unweighted statistical binning, and weighted statistical binning. The input to the pipeline is a set of sequences for different loci across different species. In the traditional pipeline, a multiple sequence alignment and gene tree is computed for each locus, and then these are given to the preferred coalescent-based summary method, and a species tree is returned. In the statistical binning pipeline, the estimated gene trees are used to compute an incompatibility graph, where each vertex represents a gene, and an edge between two genes indicates that the differences between the trees for these genes is considered significant (based on the bootstrap support of the conflicting edges between the trees). The vertices of the graph are then assigned colors, based on a heuristic for balanced minimum vertex coloring, so that no edge connects two vertices of the same color. The vertices with a given color are put into a bin, and the sequence alignments for the genes in a bin are combined into a supergene alignment. A (supergene) tree is then computed for each supergene alignment. In the unweighted binning approach (presented in [18]), these supergene trees are then given to the preferred summary method, and a species tree is returned. In the weighted binning approach presented here, each supergene tree is repeated as many times as the number of genes in its bin, and this larger set is then given to the preferred summary method.

Once the vertex coloring is computed, the genes in a given color class form a bin, and their alignments are concatenated into a supergene alignment. Then, a maximum likelihood tree is computed (perhaps with bootstrapping) on each supergene alignment, to compute a supergene tree. These supergene trees are then used by MP-EST, or some other coalescent-based summary method, to compute a species tree. Thus, statistical binning changes the input to the coalescent-based summary method, by recalculating the gene trees. Hence, statistical binning is just the first step in a coalescent-based pipeline for species tree estimation, as shown in Figure 1.

It is not hard to see that if the bin sizes are exactly the same, and statistically consistent methods are used to estimate the supergene trees and then combine them into a species tree, then the resultant pipeline is statistically consistent under the multi-species coalescent model (see the proof of Theorem 1). However, while the statistical binning heuristic developed in [18] produced bins with very similar sizes, in general it did not produce perfectly balanced bins (i.e., the bin sizes were not exactly identical). This violation of perfectly balanced bins means that the pipeline of using statistical binning to estimate gene trees, followed by MP-EST, is not a statistically consistent method for estimating species trees.

A simple variation of the technique, which was suggested in [18], corrects this problem. We keep the first step of statistical binning the same (i.e., we compute the same incompatibility graph and then use the same heuristic for balanced minimum vertex coloring), and we compute the same set of supergene trees. However, at this point we replicate every supergene tree so that it appears as many times as the number of genes in its bin. For example, if we begin with 100 genes, and obtain 20 bins, then the original statistical binning technique would produce 20 supergene trees that would be given to MP-EST to analyze. In this modified technique, if we begin with 100 genes, we end up with 100 supergene trees (although each supergene tree can appear multiple times). We call this technique “weighted statistical binning”, and refer to the original technique proposed in [18] as “unweighted statistical binning”.

Figure 1 describes the three possible pipelines (unbinned, unweighted binned, and weighted binned) for use with a summary method. In the unbinned analysis, each gene is analyzed independently, a gene sequence alignment and tree is estimated for each gene, and then a summary method, such as MP-EST, uses the gene trees to estimate the species tree. In both the weighted and unweighted binned analyses, the gene trees and sequence alignments are also computed independently, and then the incompatibility graph is formed with one vertex for each gene. In the shown example, there are 12 genes, and so the graph has 12 vertices. The 12 vertices of the incompatibility graph are then assigned colors, with two vertices colored purple, three vertices colored green, three vertices colored red, and four vertices colored blue. Note that no two vertices of the same color have an edge between them. For each color class, the sequence alignments for the associated genes are concatenated into one long supergene alignment, and a supergene tree is computed. After this point, the weighted and unweighted binning methods have different strategies. In the unweighted binning method, exactly one copy of each supergene tree is given as input to the summary method, but in the weighted binning method multiple copies of the supergene trees are given as input. Hence, in this example, MP-EST analyzes only four supergene trees in the unweighted binning pipeline, but it analyzes 12 supergene trees in the weighted binning pipeline.

Although weighted statistical binning is a very simple method, it is statistically consistent under the multi-species coalescent, as shown in Theorem 2.

Experimental Study

This study focuses on comparing weighted and unweighted statistical binning for MP-EST analyses of MLBS gene trees (the only statistically consistent coalescent method studied in [18]), using the same simulated datasets studied in [18]. We also analyze two of the biological dataests studied in [18].

Simulated datasets make it possible to accurately quantify accuracy, since the true gene trees and species trees are known, and Mirarab *et al.* [18] explored the accuracy of estimated gene trees, gene tree distributions, species tree topologies, and species tree branch lengths on simulated datasets. We analyzed the same simulated datasets using weighted statistical binning to determine whether any differences were evident between weighted and unweighted statistical binning with respect to any of these criteria. In addition, we examine the bootstrap support on the branches of the estimated species tree, as false positive edges that have low support are not as deleterious as false positive edges with high support. The bootstrap support of estimated species trees was not studied by Mirarab *et al.* [18], and therefore, this study provides the first analysis of bootstrap support for MP-EST on these datasets, as well as of the impact of of weighted or unweighted binning on the bootstrap support values.

All these aspects of phylogenomic estimation are important, but for different reasons. Species tree topologies indicate which species are more closely related to each other than to others, and so estimating accurate species tree topologies is the most important aspect of phylogenomic estimation. However, the improvement in species tree branch length estimation is also biologically relevant, since these lengths are used to estimate dates for speciation events, and also infer the amount of ILS in the data. Bootstrap support is important, since low support branches are often ignored, but high support branches are generally assumed to be correct; hence, understanding whether a method returns high support for false positive branches (indicating incorrect relations within a tree) is particularly important. Improvements in estimating the gene tree distribution matter because the accuracy of summary methods depends on an input that captures the correct gene tree distribution. Finally, improving the accuracy of gene tree estimations can result in improved understanding of protein function [1, 32].

We compute gene trees and concatenation species trees using RAxML [33] maximum likelihood. We compute coalescent-based species trees on these datasets using MP-EST with weighted and unweighted statistical binning on MLBS gene trees. For the simulated datasets, we explore species tree accuracy with respect to the true (model) species tree topology and branch lengths, and also examine the branch support of both true positive and false positive branches. We also explore the error in the gene tree distribution estimated using binning (weighted and unweighted), compared to unbinned analyses. For the biological datasets, we compare estimated species trees to the literature for each dataset, focusing on whether the estimated species tree violates known subgroups for the phylogeny.

In the avian simulation, the markers are divided into four types: exon-like, UCE-like, intron-like, and long intron-like, with exon-like having the poorest bootstrap support, long intron-like having the best, and UCE-like and intron-like intermediate in support. These differences in support are created by modifying the sequence length, so that exon-like data have the shortest sequences and least support, and long intron-like have the longest sequences and highest support. In the mammalian simulation, we also explored the impact of phylogenetic signal by varying the sequence length for the markers to produce three levels of bootstrap support. For both avian and mammalian simulations, we explore performance by varying the number of genes, the ILS level, and the sequence length. All datasets used in this study were previously used in [18].

Results and Discussion

Gene tree distribution error on avian simulated datasets

We measure the error in estimated gene tree distributions using the deviation of triplet frequencies from the triplet frequency distribution observed by true gene trees (see Methods section for details). We express these results using a cumulative distribution over all possible triplets and all replicates; hence, if a curve for one method lies above the curve for another method, then the first method strictly improves on the second method with respect to estimating the gene tree distribution. In Figure 2(a) we show results for 1000 avian genes under default ILS levels, as we vary the sequence length (from exon-like, which are the shortest loci, to long intron-like, which are the longest). In Figure 2(b) we show results with 1000 UCE-like genes, varying the ILS level. In both cases, both weighted and unweighted binning are nearly identical. Weighted and unweighted binning also showed nearly identical gene tree distribution errors under other conditions (see supplementary materials, Figs. S1). Binning improves the accuracy of

	Avian			Mammalian	
	1K genes - 1X ILS (gene tree error)	1K genes - UCE-like (ILS level)	UCE-like - 1X ILS (# of genes)	1X ILS (gene error & #)	200 genes - 63% BS (ILS level)
	Weighted binned versus Unweighted binned				
method	0.96	0.96	0.67	0.95	0.95
variable	0.99	0.99	0.99	0.99 & 0.99	0.99
	Weighted binned versus Unbinned				
method	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**0.0002**
variable	**<0.0001**	**<0.0001**	**<0.0001**	0.09 & 0.29	0.09

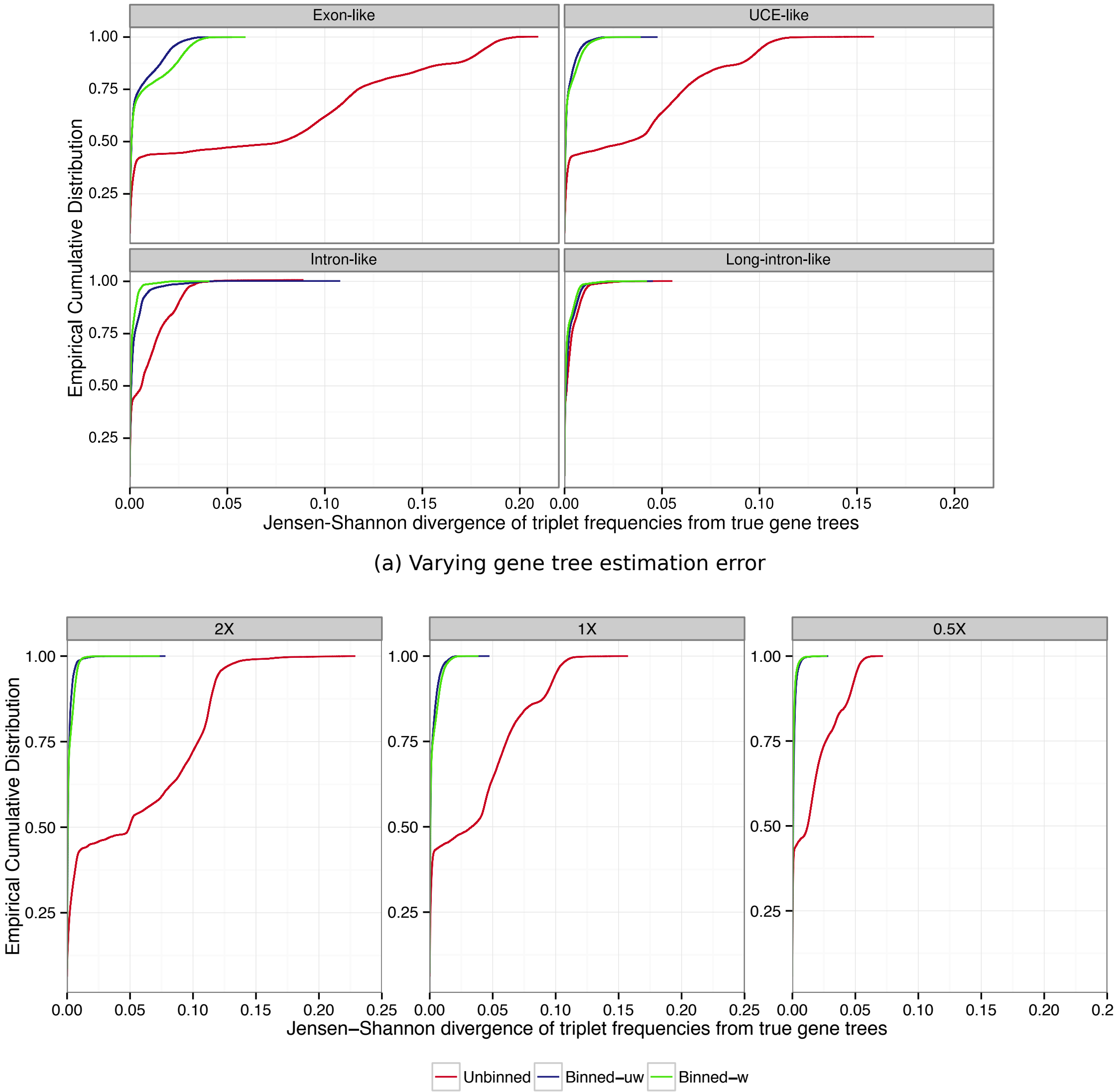
Table 1. Statistical significance test results. We performed two-way ANOVA for testing the significance of the choice of method (weighted binned versus unweighted binned on top and weighted binned versus unbinned on the bottom) and also whether there is an interaction between the choice of the method and the variable changed in each dataset (shown inside parenthesis for each column). For example, the cell under the first column and the first row titled "method" tells us that there were no statistically significant differences between the false negative rates for weighted and unweighted binned analyses on the avian dataset with 1000 genes and default ILS and varying genet tree error (p=0.96); the cell under the first column and the first row titled "variable" tells us that the gene tree error has no statistically significant impact on whether weighted or unweighted binned methods have better false negative rates (p=0.99). For the mammalian dataset, in the first column, we have two varying factors (gene tree error and the number of genes), and therefore, for interaction effects, two p-values are reported. All p-values are corrected for multiple hypothesis testing using the FDR correction (n=10 for "method" and =12 for "variable")

estimated gene tree distributions in general, but not for the longest sequences (long intron-like). Also, the improvement over unbinned analyses was highest for the lowest ILS level (2X species tree branch lengths), but was high even for the highest ILS level we explored.

Species tree estimation error on avian simulated datasets

Figure 3 shows results for species tree estimation error (FN) for analyses of avian genes of different types (exon-like through long intron-like) under the default ILS level, for 1000 UCE-like genes with varying ILS, and for varying number of UCE-like genes with default ILS. Weighted and unweighted statistical binning are essentially identical (no statistically significant differences were observed according to a two-way ANOVA test; see Table 1), and both reduce species tree estimation error compared to unbinned analyses (differences were always statistically significant with $p < 0.001$; see Table 1). The largest improvements are for the shortest gene sequences (exon-like), where unbinned analyses have approximately 23% FN rate and binned analyses have approximately 14% FN error. The difference between binned and unbinned analyses are low for intron-like sequences (11% for unbinned and 7.5% for binned), and there are no noteworthy differences for long intron-like sequences (gene tree error has a statistically significant impact; see Table 1). The impact of binning is also significantly impacted by ILS levels (see Table 1) with the largest improvements obtained for lower levels of ILS. When the number of genes is changed (see Fig. 3(c)), the impact of binning ranges from neutral to highly positive, and the largest improvements are for datasets with large numbers of genes.

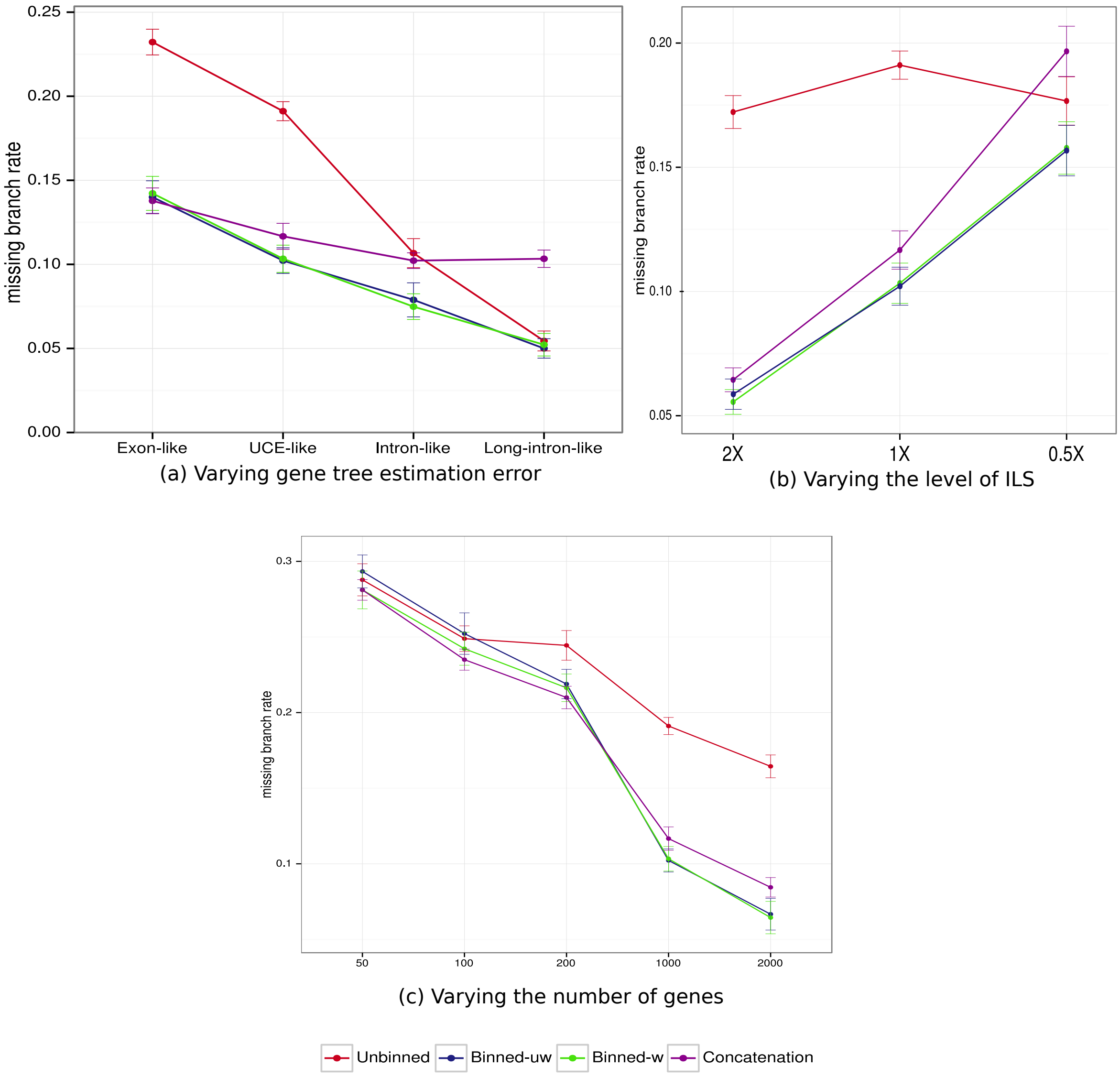
Figure 4 shows the impact of binning on species tree branch length estimation error; Figure 4(a) shows results on 1000 genes under default (1X) ILS levels and varying gene tree estimation error, and Figure 4(b) shows results on 1000 UCE-like genes with varying ILS levels. Branch length estimation accuracy is reported using the ratio of the estimated branch length to the true branch length, for those true branches recovered by the method. Thus, values equal to 1 indicate perfect accuracy, values below 1 indicate under-estimation of branch lengths (and hence over-estimation of ILS), and values above 1 indicate over-estimation of branch lengths (and hence under-estimation of ILS). Both types of binning (weighted and unweighted) produce nearly identical results with respect to branch lengths (with a slight advantage for weighted analyses). Unbinned analyses substantially under-estimate branch lengths, but as



(a) Varying gene tree estimation error

(b) Varying the ILS level

Figure 2. Divergence of estimated gene tree (triplet) distributions from true gene tree distributions for MP-EST analyses of simulated avian datasets. In (a), we vary the gene tree estimation error (exon-like has the highest error, and long intron-like has the lowest error) and explore 1000 genes under default ILS levels, and in (b) we vary the amount of ILS and fix the number of genes to 1000 and sequence length to 500bp (UCE-like genes). 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 from estimated gene/supergene trees. For each of these $\binom{n}{3}$ triplets, we calculate the Jensen-Shannon divergence of the estimated triplet distribution from the true gene tree triplet distribution. We show the empirical cumulative distribution of these divergences. The empirical cumulative distribution shows that for a given divergence level, the percentage of the triplets that are diverged from true triplet distribution at or below that level. Results are shown for 10 replicates.



(a) Varying gene tree estimation error

(b) Varying the level of ILS

(c) Varying the number of genes

Figure 3. Species tree estimation error (FN) for MP-EST with MLBS on avian simulated datasets. We (a) vary the gene tree estimation error (exon-like genes have the highest error, and long intron-like have the lowest error) and fix the number of genes to 1000 with default amount of ILS (1X level), (b) vary the ILS level and fix the number of genes to 1000 and the sequence length to UCE-like, and (c) vary the number of genes with fixed default level of ILS (1X level) and UCE-like sequence length. We show results for 20 replicates everywhere, except for 2000 genes that are based on 10 replicates.

the sequence length increases, the branch length estimations produced by unbinned analyses improve, so that they are more accurate at the long intron-like markers. Using binning (either type) improves branch length estimation, and the improvement is very large for the shorter sequences (exon-like and UCE-like). When levels of ILS are changed, weighted and unweighted binning are again close (with a

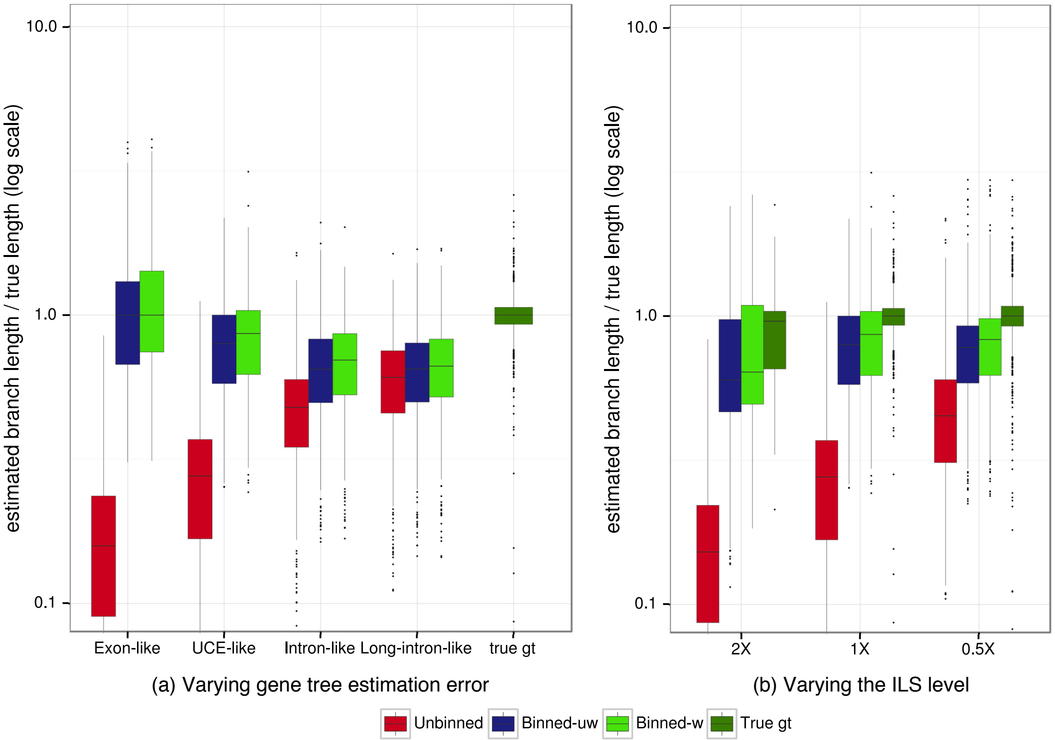
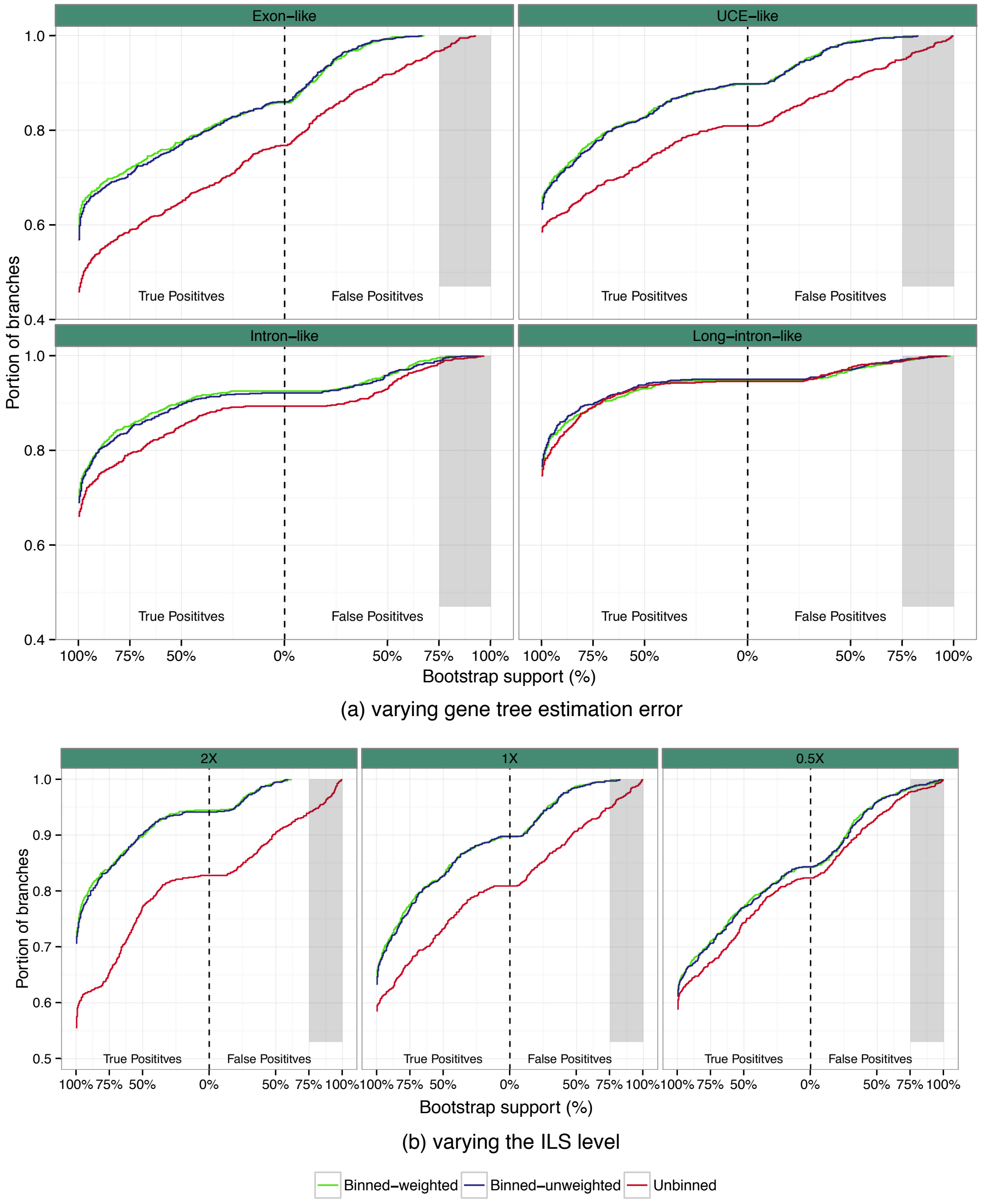


Figure 4. Effect of binning on the species tree branch lengths (in coalescent units) estimated by MP-EST using MLBS on the avian simulated datasets. In (a) we vary gene estimation error, using 1000 genes with default ILS (1X branch length), and in (b) we vary the ILS level using 1000 UCE-like genes. We show the species tree branch length accuracy (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, values below 1 indicate under-estimation of branch lengths, and values above 1 indicate over-estimation of branch lengths.

slight advantage for weighted), and show little change in branch length estimation with changes in ILS levels; however, unbinned analyses substantially under-estimate branch lengths for the lowest ILS model condition, and then become more accurate (although still under-estimate) with increases in the ILS level. Hence, the biggest improvement obtained by binning is for the lowest ILS (2X branch lengths), and there is less improvement for the highest ILS level (0.5X). The likely explanation for this trend is that MP-EST interprets all discord as due to ILS, and produces a model tree (with branch lengths) that it considers most likely to generate the observed discordance. Hence, MP-EST tree branch lengths will be closer to the correct lengths when the ILS level is very high.

Bootstrap support on avian simulated datasets

We explore bootstrap support of trees estimated on simulated avian datasets, as follows. We assign relative quality to each edge in an estimated tree, taking bootstrap support into account. The highest quality edges are the true positive branches with the highest bootstrap support, and the lowest quality edges are the false positive branches with the highest bootstrap support, and all other edges fall in between. We



(a) varying gene tree estimation error

(b) varying the ILS level

Figure 5. Cumulative distribution of the bootstrap support values (obtained using MLBS) of true positive (TP) and false positive (FP) edges estimated by MP-EST on mammalian datasets. In (a) we fix the number of genes to 1000, use default ILS levels, and vary gene tree estimation error (exon-like have the highest error, and long intron-like have the lowest error), and in (b) we study 1000 UCE-like genes and vary ILS levels. To produce the graph, we order the branches in the estimated species tree by their quality, so that the true positives with high support come first, followed by lower support true positives, then by false positives with low support, and finally by false positives with high support. The false positive branches with support above 75% are the most troublesome, and that fraction are indicated in the grey area. When the curve for a method lies above the curve for another method, then the first method has better bootstrap support.

order all the edges by their quality, so that the true positive branches come first (with the high support branches before low support branches), followed by the false positive branches (with the low support branches before the high support branches). Given this ordering, we create the subfigure, where the x-axis indicates the edge quality (from very high to very low, as you move from left to right), and the y-axis indicates the fraction of the edges having at least the quality indicated by the x-axis. Thus, the higher the curve, the better the overall quality of the species tree.

Figure 5 shows results on 1000 avian genes (20 replicates) under default ILS and with varying sequence length, and also with 1000 UCE-like genes with varying ILS levels. Both types of binning are nearly identical in terms of their impact on bootstrap support, and both improve bootstrap support; in particular, using binning increases the number of highly supported true positive branches and decreases the number of highly supported false positives. However, the sequence length modulates the impact of binning on bootstrap support, so that the largest impact is for the shortest sequences (exon-like) and there is insignificant impact for the longest sequences (long intron-like). ILS levels also impacts how binning affects the bootstrap support, so that the biggest improvement in bootstrap support is obtained for the lowest ILS level (2X branch lengths) The number of genes also impacts the bootstrap support (Fig. S2) so that the biggest improvement in bootstrap support is obtained for the largest number of genes (2000) (and there is little to no difference between binned and unbinned analyses on 50 or 100 genes); furthermore, weighted and unweighted binning produce very similar bootstrap supports.

Comparisons to concatenation on avian simulated datasets

We compare both weighted and unweighted binning analyses to unbinned analyses, as well as to concatenation using maximum likelihood on the avian datasets. Weighted and unweighted binning with MP-EST have essentially identical accuracy on 1000 genes simulated under default ILS levels (Fig. 3(a)). Concatenation matches the accuracy of these binned MP-EST methods on the shortest sequences (exon-like), but then is less accurate than them for longer sequences. Unbinned analyses are much less accurate than both concatenation and binned analyses, except at the longer sequences (where it matches concatenation at intron-like sequences, and then improves on concatenation and matches the binned analyses for the long intron-like sequences). Weighted and unweighted binned MP-EST trees are again nearly identical when we vary the levels of ILS (Fig. 3(b)) or the number of genes (Fig. 3(c)). Both binned analyses are more accurate than concatenation and unbinned analyses at all ILS levels (Fig. 3(b)), but with few genes, concatenation can be slightly more accurate than binned or unbinned analyses. Thus, compared to concatenation, binned analyses have their highest advantage for longer gene sequences, lower ILS levels, and higher number of genes.

Comparisons on simulated mammalian datasets

Results on simulated mammalian datasets are similar to analyses of avian datasets. In nearly every condition, both weighted and unweighted binning show very similar results. Binning improved gene tree distributions, generally with very large improvements, and the improvements decreased with the sequence length and ILS level (Fig. S3). Binning also tended to improve species tree topology estimation (Fig. 6), but the impact depended on the sequence length (binning is beneficial for shorter sequences and neutral for longer sequences) and number of genes (binning can dramatically improve species tree topologies given a large number of genes, but can be neutral or even detrimental for a small number of genes). ILS level also has an impact, so that binning is most helpful for low ILS levels, and less helpful for high ILS levels (Fig. S4).

As observed in the avian simulations, unbinned analyses substantially under-estimated species tree branch lengths. Both weighted and unweighted binning produced nearly identical branch lengths for all sequence lengths, number of genes, and ILS levels, and both types of binning came closer to the true branch lengths than unbinned analyses (Figs. 7 and S5).

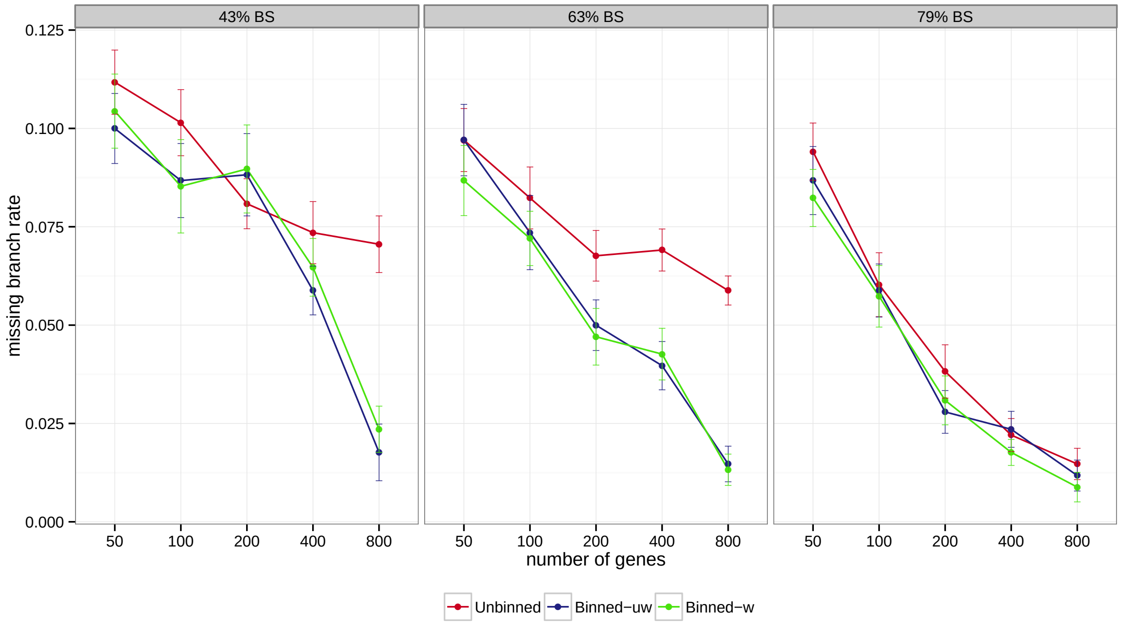


Figure 6. Species tree estimation error for MP-EST with MLBS on mammalian simulated datasets. We show average FN rate over 20 replicates. We varied the number of genes (50, 100, 200, 400 and 800) and sequence length (250bp (43% BS), 500bp (63% BS) and 1000bp (79% BS)) with default amount of ILS (1X level).

Finally, both weighted and unweighted binning produced nearly identical species tree branch support values, and both matched or improved unbinned analyses for all tested numbers of genes, sequence lengths, and ILS levels (Figs. S6 and S7). However, improvements increased with the number of genes and decreased with the sequence length and ILS level.

Analysis of biological datasets

We compared weighted and unweighted binning of MP-EST on MLBS gene trees on two biological datasets studied in [18] – the avian dataset studied in [21] and a mammalian dataset (obtained by deleting 23 erroneous genes from the dataset studied in [34]). There were no topological differences between MP-EST trees estimated using weighted or unweighted statistical binning, and extremely small differences in branch support (less than 3%; see Fig. 8). Thus, although [21] used unweighted statistical binning, the main conclusions they drew about the evolutionary history of modern birds are also found in the weighted statistical binning of both types of gene trees. Note that the unbinned analysis violates several subgroups established in the avian phylogenomics project and other studies (indicated in red in Fig. 8), but that the binned analysis does not violate any of these subgroups. Of these violated subgroups, the failure of the unbinned analysis to recover Australaves is the most significant, since it has been recovered in many prior analyses [23–26].

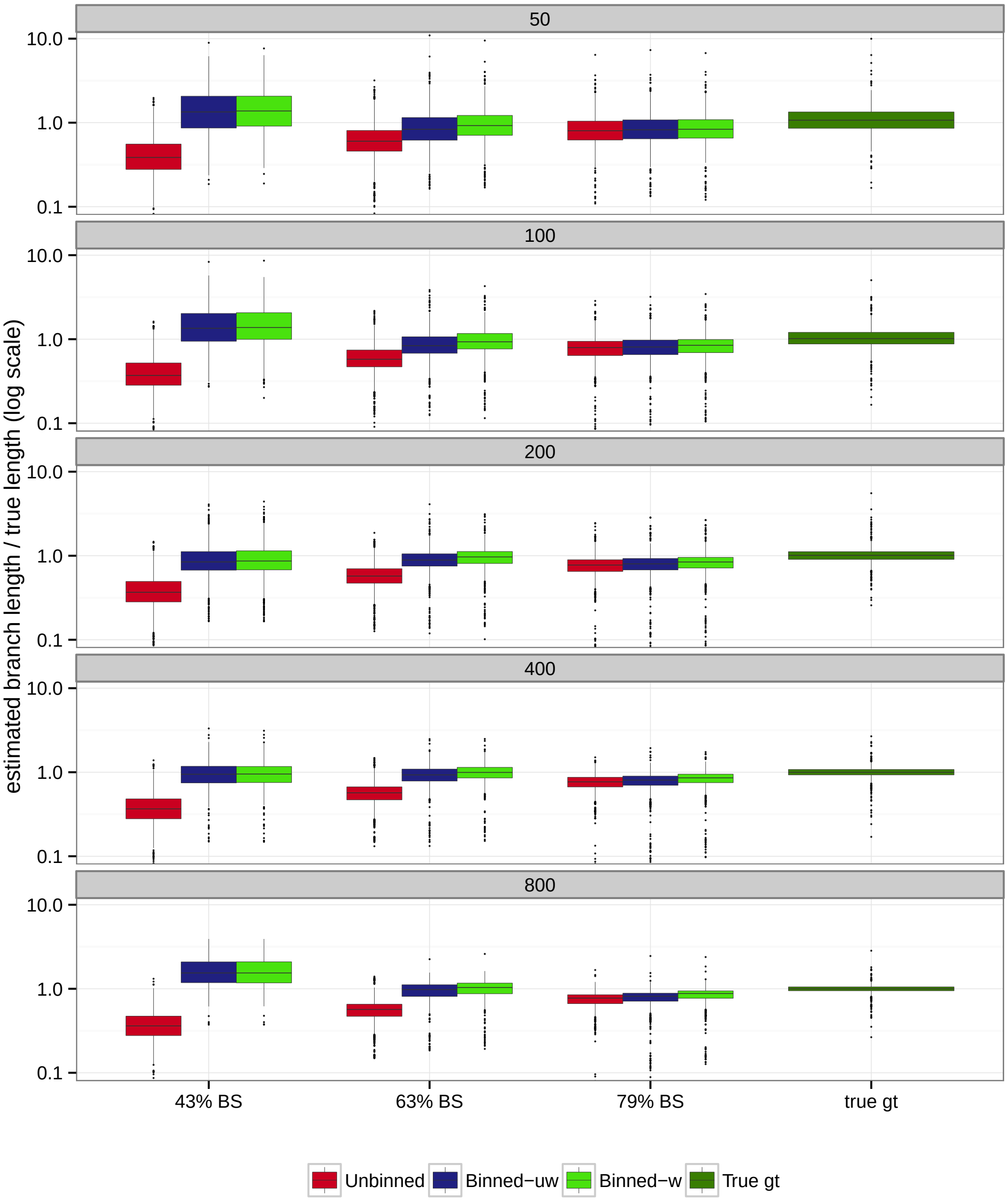


Figure 7. Effect of binning on the branch lengths (in coalescent units) estimated by MP-EST using MLBS on the mammalian simulated datasets. We show the species tree branch length error (the ratio of estimated branch length to true branch length for branches of the true tree that appear in the estimated tree; 1 indicates correct estimation). We varied the number of genes (50, 100, 200, 400 and 800) and sequence length (250bp (43% BS), 500bp (63% BS) and 1000bp (79% BS)) with default amount of ILS (1X level).

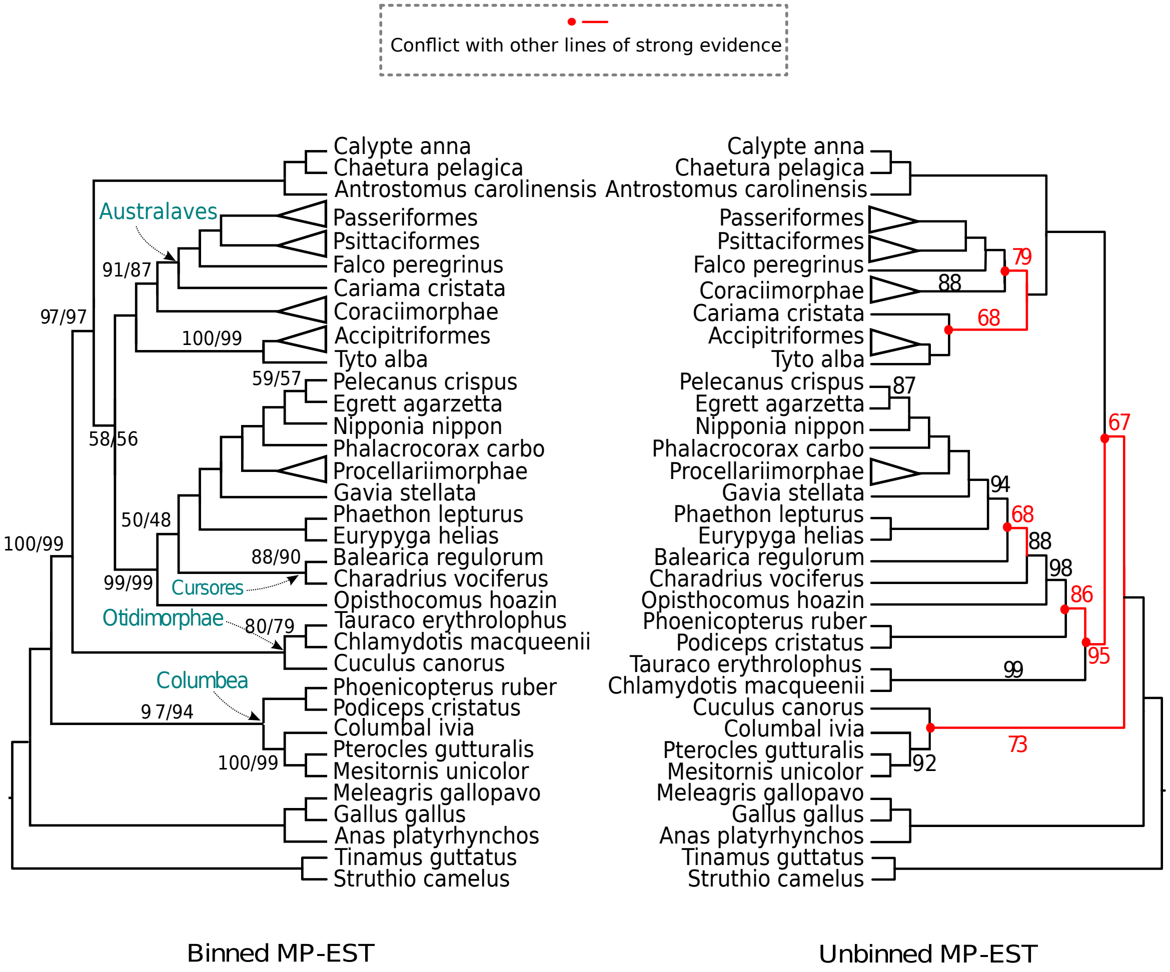


Figure 8. Trees computed on the avian biological dataset using MP-EST on MLBS gene trees. We show results with weighted and unweighted binning (left), and unbinned analyses (right). MP-EST with weighted and unweighted binning returned the same tree. The branches on the binned MP-EST tree are labeled with two support values side by side: the first is for unweighted binning and the second is for weighted binning; branches without designation have 100% support. Branches in red indicate contradictions to known subgroups.

Conclusions

Because species trees and gene trees can differ, the estimation of species trees requires multiple loci. One approach to estimating species trees from multiple conflicting loci seeks to restrict the set of loci using principled arguments [9], but other approaches that explicitly model the discordance have also been developed. When gene tree discord is due to incomplete lineage sorting, then statistical methods, such as

MP-EST, can be used to estimate the species tree by combining gene trees. However, this study, as well as others [16–19, 35, 36], demonstrates that gene tree estimation error impacts species tree estimation, so that species trees estimated using summary methods on poorly estimated gene trees can have low accuracy. The (unweighted) statistical binning technique proposed in [18] improved the accuracy of estimated gene trees, and was shown to improve the accuracy of MP-EST when applied to MLBS gene trees. However, the use of unweighted statistical binning to estimate a species tree does not have the desirable property of being statistically consistent. This is a significant drawback for unweighted statistical binning, which could discourage biologists from using the technique.

This study described a simple modification to statistical binning, obtained by replicating each supergene tree by the number of genes in its bin (equivalently, replacing each gene tree from the input set by its recalculated tree, which is the supergene tree for the bin). This modification, which we call "weighted statistical binning", is statistically consistent under the multi-species coalescent, and so addresses this drawback. Also, weighted statistical binning produced nearly identical results to unweighted statistical binning on the simulated datasets, and topologically identical results (with very similar bootstrap support values) on the biological datasets we explored in this study, and so this study supports the findings about the avian phylogeny reported in [21]. The fact that weighted and unweighted binning produced similar results is not surprising, since that unweighted binning technique strived to create "balanced" bins as much as possible, and largely achieved this on the datasets we explored. Thus, in practice, weighted and unweighted binning may generally have very similar performance on many datasets.

Under the model conditions we studied, both weighted and unweighted statistical binning generally improve gene tree distributions, species tree topologies and branch lengths, and bootstrap support (so that statistical binning increases bootstrap support for true positive edges, and reduces the number of highly supported false positives), compared to unbinned analyses. These improvements increase when gene sequence alignments have low phylogenetic signal, the species tree has low ILS, or there are many genes.

Although most comparative studies have focused on the estimation of the species tree topology, the other aspects of phylogenomic estimation are also important. For example, the improvement in species tree branch length is biologically significant since these lengths are used to estimate dates for speciation events, and to infer the amount of ILS in the data. Since unbinned analyses tended to substantially under-estimate branch lengths, this means that MP-EST used without binning tends to over-estimate ILS. Binning largely corrects this and produces branch lengths that are much closer to their true lengths. Since MP-EST tends to over-estimate ILS in the presence of gene tree estimation error, this means that predictions of ILS levels for biological datasets may have over-estimated these amounts. The improvement in branch support is also biologically relevant, especially since binning generally increased the number of strongly supported true positive branches and reduced the number of strongly supported false positive branches. Finally, improvements in gene tree estimation could lead to improved understanding of protein structure and function [1, 32].

Results on the two biological datasets show that weighted and unweighted statistical binning analyses produced identical species trees and nearly identical branch support values; furthermore, these binned analyses were more congruent with established subgroups than unbinned analyses. Thus, the analyses of biological datasets supports the trends on simulated datasets showing that weighted statistical binning is beneficial for species tree estimation using MP-EST on MLBS gene trees.

Despite the positive impact of weighted statistical binning on species tree estimation in this study, new research published in the last few months suggests that these techniques need to be evaluated more thoroughly, using other datasets and other ways of computing species trees from multiple gene trees.

For example, Mirarab *et al.* [37] showed that the accuracy of MP-EST trees depended on whether MLBS or best maximum likelihood (BestML) gene trees were used, and that MP-EST trees based on BestML gene trees generally produced more accurate species tree topologies for datasets with large numbers of genes (such as the ones we studied in this paper). The explanation offered for this is that

BestML gene trees are generally more accurate than MLBS gene trees, and that this helps coalescent-based species tree estimation. Hence, the evaluation of the impact of binning on MP-EST with BestML gene trees is also needed. It is also possible that better results would be obtained using Bayesian methods (such as MrBayes [38]), rather than MLBS, to generate the distribution of gene trees [39], since the posterior distribution produced by Bayesian MCMC methods may not have as high gene tree estimation error as the MLBS sample. Similarly, methods that co-estimate alignments and trees, such as BAli-Phy [40], SATé [41, 42], and PASTA [43], might provide improved accuracy compared to the standard two-step procedure for estimating trees (first align, and then compute the tree).

Another important limititation of this study, as well as of [18], is that it only explored the impact of binning on MP-EST, and not on any other statistically consistent coalescent-based method; yet, ASTRAL [36] is a new statistically consistent coalescent-based method that provides better accuracy than MP-EST under many conditions. One possible explanation for why ASTRAL is more accurate than MP-EST is that it is more robust to gene tree estimation error. However, to the extent that the ASTRAL achieves good accuracy due to increased robustness to gene tree estimation error, binning may not be beneficial, and could even be hurtful. Hence, statistical binning needs to be tested on other coalescent-based summary methods, especially ones like ASTRAL that may be more robust to gene tree estimation error. ASTRAL can also analyze large datasets (such as the plant transcriptome dataset with approximately 100 species and 800 loci [44]), and unlike MP-EST does not require rooted gene trees; hence, understanding the impact of binning on ASTRAL's accuracy is also of practical importance.

Finally, although this study and [18] explored a large number of datasets, these studies were fairly limited in terms of the types of datasets they examined. For example, neither study looked at the impact of binning for very small numbers of species (e.g., below 20), nor for small numbers of genes (e.g., below 50). Neither study examined model conditions in which gene tree estimation error is due to model misspecfication, nor where the different loci evolve under different sequence evolution models (e.g., different GTR matrices). Neither study considered biological causes for gene tree discord, such as gene duplication and loss or horizontal gene transfer, other than ILS.

Therefore, while this study shows the beneficial impact of weighted statistical binning on species tree estimation, the benefit has only been demonstrated for one coalescent-based method (MP-EST on MLBS gene trees) and only under a limited set of model conditions. It is possible that the benefits seen for weighted statistical binning may not hold when used with other coalescent-based methods, or under other model conditions. Clearly further investigation is needed to determine the conditions where weighted statistical binning improves the estimation of gene trees and species trees for multi-locus phylogenomic analysis, and whether there are conditions where weighted statistical binning can be hurtful.

Indeed, one interpretation of this study is that it suggests that the research focus should be directed at developing more accurate methods for gene tree estimation, as well as developing new coalescent-based methods that are highly robust to gene tree estimation error. It is also possible that methods that construct species trees directly from the sequence data, rather than by combining gene trees, will have the best accuracy (see, for example, [45–47]).

Despite these caveats, this study confirms the general finding in [18] that highly accurate coalescent-based species tree estimation is possible, and that good coalescent-based methods can provide improved accuracy relative to concatenation under many biologically realistic conditions, even for genome-scale datasets with short sequences on each locus, where there are high amounts of gene tree discord due to incomplete lineage sorting. However, as observed in this and other studies [19, 35], concatenation often produces more accurate trees than even the best coalescent-based methods when the level of ILS is low enough. Therefore, an important question is whether a given biological dataset has a sufficiently high level of ILS that a coalescent-based analysis is needed. Conversely, coalescent-based methods that are not only more accurate than concatenation under conditions with high ILS, but also comparably accurate even under low levels of ILS, would be very helpful tools.

Materials and Methods

Theoretical results

We begin with a negative result, showing that unweighted statistical binning is not statistically consistent under the multi-species coalescent model. We consider phylogenomic pipelines that begin with sequence alignments, compute gene trees on the alignments using statistically consistent methods (such as maximum likelihood), and then combine the gene trees using statistically consistent methods (such as MP-EST) that only consider gene tree topologies and not also their branch lengths. Thus, the theoretical results apply to statistically consistent coalescent-based summary methods such as MP-EST, the population tree from BUCKy, and ASTRAL. We add that there are no mathematical proofs yet of statistical consistency for summary methods, when gene trees have any estimation error. Hence, our proofs will also assume that sequence lengths are long enough that all gene trees can be estimated completely correctly, with high probability.

Theorem 1 *Phylogenomic pipelines using unweighted statistical binning, a statistically consistent gene tree estimation method, and a statistically consistent summary method, are not statistically consistent under the multi-species coalescent model.*

Proof: Let $t < 1$ be the threshold support provided by the user, let T be the true species tree, and let $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$ be a set of true gene trees generated by T under the multi-species coalescent model. Let A_i be a set of sequences generated on t_i under a sequence evolution model (such as the Generalized Time Reversible Model [48]). Then, given long enough sequences, if an appropriate method is used (such as maximum likelihood), then with high probability the gene trees estimated on each A_i will be topologically correct and all branches will have support at least t. Hence, when analyzed using weighted statistical binning, no two genes with topologically different trees will be placed in the same bin. Consequently, each bin will have a set of genes, and the true gene trees for the genes in the bin will have the same topology. If a partitioned analysis (allowing all model parameters other than the tree topology) is performed on the concatenated alignments of these genes, and the method used is statistically consistent, then the tree computed on the bin will have the same topology as the true gene tree for the bin. For a large enough number of genes, every possible tree on the n leaves will be the true gene tree for at least one gene with high probability. When this happens in an unweighted statistical binning analysis, the set of gene trees given as input to the coalescent-based summary method will contain each possible gene tree exactly once – i.e., the estimated distribution on gene tree topologies will be flat. No summary method can infer the correct species tree from a flat distribution, and so the pipeline will be statistically inconsistent. $\square$

Note that the negative result does not imply that the pipeline is positively misleading (i.e., that the pipeline would return an incorrect tree with high confidence), but rather that it would not be able to converge to any tree. Thus, unweighted statistical binning becomes *uninformative*, as the number of genes increases. We now prove the second theoretical result, which shows that weighted statistical binning makes these pipelines statistically consistent.

Theorem 2 *Phylogenomic pipelines using weighted statistical binning, a statistically consistent gene tree estimation method, and a statistically consistent summary method, are statistically consistent under the multi-species coalescent model.*

Proof: The proof is nearly identical to that of Theorem 1, except that we prove statistical consistency because the binning is weighted. Let $t < 1$ be the threshold support provided by the user, let T be the true species tree, and let $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$ be a set of true gene trees generated by T under the multi-species coalescent model. Let A_i be a set of sequences generated on t_i under a sequence evolution model (such as the Generalized Time Reversible Model [48]). Then, given long enough sequences, if an appropriate method is used (such as maximum likelihood), then with high probability the gene trees estimated on each A_i will be topologically correct and all branches will have support at least t. Hence, when analyzed using weighted statistical binning, no two genes with topologically different trees will be placed in the same bin. Consequently, each bin will have a set of genes, and the true gene trees for the genes in the bin will have the same topology. If a partitioned analysis is performed on the concatenated alignments of these genes, and the method used is statistically consistent, then the tree computed on the bin will have the same topology as the true gene tree for the bin. Therefore, as the number of genes increases, the gene tree distribution defined by the weighted binning technique will converge to the true gene tree distribution. Hence, for any statistically consistent coalescent-based method, the species tree estimated on these supergene trees will converge to the true species tree. □

Evaluation

We explored the performance of MP-EST and concatenation (using RAxML) using weighted and unweighted binning on a collection of simulated and biological datasets originally studied in [18]. We applied MP-EST to a set of RAxML gene trees computed on bootstrap replicates of each gene sequence alignment. With bootstrap ML gene trees for each gene, summary methods were applied with the site-only multi-locus bootstrapping (MLBS) procedure [22], implemented as follows. For each gene or supergene, 200 replicates of bootstrapping are performed using RAxML. Next, 200 replicates $(R_1, R_2, \ldots, R_{200})$ of input datasets to the summary methods are created such that R_i contains the i^{th} bootstrap tree across all genes/supergenes. The summary methods are then run on these 200 input replicates, and 200 species trees are estimated. Finally, the greedy consensus tree of these 200 estimated species tree is computed, and support values are drawn on the branches of the greedy consensus tree by counting the occurrences of each bipartition in the 200 species trees.

Gene tree distribution error

MP-EST computes species trees using the estimated distribution on rooted triplet trees defined by its input of gene trees. We therefore evaluated the impact of binning on the estimated gene tree distribution, measuring the divergence between the triplet distribution of estimated gene trees and the triplet distribution of true gene trees. We represent the gene tree distribution 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. Therefore, we have $\binom{n}{3}$ true triplet distributions. Hence, for each triplet of taxa, we have estimated triplet distributions using the unbinned analysis, as well as weighted and unweighted binning analyses. We computed the Jensen-Shannon divergence of each of these $\binom{n}{3}$ triplet distributions and showed the empirical cumulative distribution of these divergences. The Jensen-Shanon divergence is a symmetrized and smoothed version of Kullback-Leibler divergence [49] between two distributions P and Q, and can be calculated as follows [50]:

$$JS(P, Q) = \frac{1}{2}KL(P, M) + \frac{1}{2}KL(Q, M) \qquad (1)$$

where $M = \frac{P+Q}{2}$, and KL is the Kullback-Leibler divergence.

Species tree estimation error and branch support

We compared the estimated species trees to the model (i.e., true) species tree (for the simulated datasets) or to the scientific literature (for the biological datasets). We measure topological error using the missing branch rate (also known as the false negative (FN) rate), which is the proportion of branches in the true tree that are missing from the estimated tree. We also reported the error in species tree branch lengths estimated by MP-EST using the ratio of estimated branch length to true branch length for those branches of the true tree that appear in the estimated tree; thus, 1 indicates correct estimation, values above 1 indicate lengths that are too long, and values below 1 indicate branch lengths that are too short. Note that species tree branch lengths reflect the expected amount of ILS, and so under-estimation of species tree branch lengths means over-estimation of ILS, and over-estimation of branch lengths means under-estimation of ILS. We also computed the branch support of the false positive (FP) and true positive (TP) edges, where false positive edges are present in the estimated tree but not in the true tree, and edges that are present in both the estimated and true tree are true positive edges.

Simulated datasets

We studied simulated datasets based on the mammalian dataset of [34] and the avian dataset of [21], which were generated in a prior study [18]. We briefly describe the simulation protocol for the biological datasets, and direct the reader to [18] for full details.

Mammalian simulated datasets

This dataset was generated by [18], and studied there and also in [37]. Here we describe the procedure followed by [18] to generate these data. First, a species tree was computed for the full biological dataset in [34], using MP-EST, and the tree topology and branch lengths were used as the model tree. Thus, the mammalian simulation model tree has an ILS level based on an MP-EST analysis of the biological mammalian dataset. Gene trees were simulated within this species tree under the multi-species coalescent model, and then the branch lengths on the gene trees were defined using the gene trees estimated on the biological dataset.

Variants of the basic model condition were generated by varying the amount of ILS, the number of genes, and the sequence length for each gene; these modifications also impact the amount of gene tree estimation error and the average bootstrap support in the estimated gene trees, and so can be modified to produce datasets that resemble the biological data.

The amount of ILS was varied by adjusting the branch length (shorter branches increase ILS). A model condition with reduced ILS was created by uniformly doubling (2X) the branch lengths, and a model condition with higher ILS was generated by uniformly dividing the branch lengths by two (0.5X). The amount of ILS obtained without adjusting the branch lengths is referred to as “default ILS”, and was estimated by MP-EST on the biological data.

The average bootstrap support (BS) in the biological data was 71%, and so [18] generated sequence lengths that produced estimated gene trees with bootstrap support bracketing that value – 500bp alignments produced estimated gene trees with 63% average BS and 1000bp alignments produced estimated gene trees with 79% BS. They also generated model conditions with very short sequence length (250bp), which has 43% average BS.

Mirarab et al. [18] varied the number of genes from 50 to 800 to explore both smaller and larger numbers of genes than the biological dataset (which had roughly 400 genes). In total, [18] generated 17 different model conditions specified by the ILS level, the number of genes, and the sequence length. For each of these model conditions, [18] created 20 replicates, with the exception of the two model conditions with 400 and 800 genes of 250bp sequence length, where they had 10 and 5 replicates, respectively.

Avian simulated datasets

Mirarab et al. [18] used the species tree estimated by MP-EST on the avian dataset with 48 species and 14,446 loci studied by [21], and simulated gene trees by varying different parameters (similar to the mammalian simulated datasets). Three types of genomic markers were studied: exons, UCEs, and introns. The average bootstrap support (BS) of the gene trees based on exons, UCEs, and introns, was 24%, 39% and 48%, respectively. The longest introns had the highest average BS (59%). Mirarab et al. generated four model conditions that resemble these four markers, and refer to these model conditions as exon-like, UCE-like, intron-like and long-intron-like. Mirarab et al. varied the number of genes from 50 to 2000 and the amount of ILS, using the same technique as the mammalian simulated datasets.

Biological datasets

We studied two biological datasets also studied in [18]: the avian dataset [21] containing 14,446 loci across 48 species, and a reduced version of the mammalian dataset studied by Song et al. [34] with 447 loci across 37 species, from which [18] deleted 23 erroneous genes (see [18, 37] for discussion of these loci).

Methods and commands

Gene tree estimation: RAxML version 7.3.5 [51] was used to estimate gene trees under the GTRGAMMA model, using the following command:

```
    raxmlHPC-SSE3 -m GTRGAMMA -s [input_alignment] -n [output_name] -N 20
-p [random_seed_number]
```

The following command was used for bootstrapping:

```
    raxmlHPC-SSE3 -m GTRGAMMA -s [input_alignment] -n [output_name] -N 200
-p [random_seed_number] -b [random_seed_number]
```

Concatenation: For the concatenated analysis, we computed a parsimony starting tree using RAxML version 7.3.5, and then ran RAxML-light version 1.0.6. We used the following commands:

```
    raxmlHPC-SSE3 -y -s supermatrix.phylip -m GTRGAMMA -n [output_name]
-p [random_seed_number]
    raxmlLight-PTHREADS -T 4 -s supermatrix.phylip -m GTRGAMMA -n name
-t [parsimony_tree]
```

MP-EST: We used version 1.3 of MP-EST. We ran MP-EST 10 times with different random seed numbers, and selected the species tree with the best likelihood score using a custom shell script. MP-EST was run using site-only multi-locus bootstrapping, using 200 MLBS replicates, and returning the greedy consensus of the 200 MP-EST MLBS species trees as the output. The branch support on the edges of the tree represent the frequency of the bipartition in the sample of 200 species trees.

Greedy consensus: The greedy consensus (also called the "extended majority consensus") of a set of trees, all on the same set of leaves, is obtained by ordering the bipartitions that appear in one or more trees in the order of their frequency (most frequent first). Then, a tree is built from this set, beginning with the first bipartition, and then modifying the tree to include the next bipartition in the list, if the addition of the bipartition is possible. We used Dendropy version 3.12.0 [52] to compute greedy consensus trees when running MP-EST with MLBS gene trees.

Acknowledgements

The authors acknowledge the support from the Texas Advanced Computing Center (TACC), which provided a generous allocation of computing time so that these analyses could be performed. This work was supported by National Science Foundation (NSF) grant numbers 1461364, 0733029, 1062335 to TW, MSB, and SM, and also by Howard Hughes Medical Institutes (HHMI) graduate international student fellowship to SM. All data are available on `doi.org/2142/55319` and the code can be found at `https://github.com/smirarab/binning`.

References

1. Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Research 8: 163–167.
2. Maddison W (1997) Gene trees in species trees. Systematic Biology 46: 523–536.
3. Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecology Evolution 26.
4. Edwards SV (2009) Is a new and general theory of molecular systematics emerging? Evolution 63: 1–19.
5. Rosenberg NA (2013) Discordance of species trees with their most likely gene trees: A unifying principle. Molecular Biology and Evolution 30: 2709-2713.
6. Roch S, Steel M (2014). Likelihood-based tree reconstruction on a concatenation of alignments can be positively misleading. Arxiv publication arXiv:1409.2051.
7. Degnan JH, DeGiorgio M, Bryant D, Rosenberg NA (2009) Properties of consensus methods for inferring species trees from gene trees. Syst Biol 58: 35–54.
8. Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology 56: 17–24.
9. Salichos L, Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497: 327–31.
10. Wu Y (2012) Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution 66: 763-775.
11. Kubatko LS, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinfomatics 25: 971–973.
12. Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution 27: 570–580.
13. DeGiorgio M, Degnan JH (2010) Fast and consistent estimation of species trees using supermatrix rooted triples. Molecular Biology and Evolution 27: 552–569.
14. Liu L, Yu L, Pearl DK, Edwards SV (2009) Estimating species phylogenies using coalescence times among sequences. Systematic Biology 58: 468–477.
15. Liu L, Yu L, Edwards S (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evolutionary Biology 10: 302.
16. Bayzid M, Warnow T (2013) Naive binning improves phylogenomic analyses. Bioinformatics 29: 2277–84.
17. Patel S, Kimball RT, Braun EL (2013) Error in phylogenetic estimation for bushes in the tree of life. Journal of Phylogenetics and Evolutionary Biology 1: 110.
18. Mirarab S, Bayzid MS, Boussau B, Warnow T (2014) Statistical binning improves species tree estimation in the presence of gene tree incongruence. Science 346: 1250463.
19. Gatesy J, Springer M (2014) Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Molecular Phylogenetics and Evolution 80: 231-266.
20. Lanier H, Knowles L (2012) Is recombination a problem for species-tree analyses? Syst Biol 61: 691–701.
21. Jarvis ED, Mirarab S, et al. (2014) Whole genome analyses resolve early branches in the tree of life of modern birds. Science 346: 1320-1331.
22. Seo TK (2008) Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Molecular biology and evolution 25: 960–971.
23. Suh A, Paus M, Kiefmann M, Churakov G, Franke FA, et al. (2011) Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. Nature Comm 2.
24. Wang N, Braun E, Kimball R (2012) Testing hypotheses about the sister group of the Passeriformes using an independent 30-locus data set. Mol Biol Evol 29: 737-750.

25. Kimball RT, Wang N, Heimer-McGinn V, Ferguson C, Braun EL (2013) Identifying localized biases in large datasets: A case study using the avian tree of life. Molecular Phylogenetics and Evolution 69: 1021-1032.

26. McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, et al. (2013) A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS ONE 8: e54848.

27. Gusfield D (1991) Efficient algorithms for inferring evolutionary trees. Networks 21: 19–28.

28. Warnow T (1994) Tree compatibility and inferring evolutionary history. Journal of Algorithms 16: 388-407.

29. Warnow T, Moret BM, St John K (2001) Absolute convergence: true trees from short sequences. In: Proceedings of the twelfth annual ACM-SIAM Symposium on Discrete Algorithms (SODA). pp. 186–195.

30. Karp R (1972) Reducibility among combinatorial problems. In: Miller RE, Thatcher JW, editors, Complexity of Computer Computations, Plenum. pp. 85-103.

31. Brélaz D (1979) New methods to color the vertices of a graph. Communications of the ACM 22: 251–256.

32. Sjolander K (2004) Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics 20: 170–179.

33. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.

34. Song S, Liu L, Edwards SV, Wu S (2012) Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proceedings of the National Academy of Sciences of the United States of America 109: 14942–7.

35. Leaché AD, Rannala B (2011) The accuracy of species tree estimation under simulation: a comparison of methods. Systematic Biology 60: 126–37.

36. Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson S, et al. (2014) ASTRAL: Genome-scale coalescent-based species tree estimation. Bioinformatics 30: i541-i548.

37. Mirarab S, Bayzid MS, Warnow T (2014) Evaluating summary methods for multi-locus species tree estimation in the presence of incomplete lineage sorting. Systematic Biology .

38. Huelsenbeck J, Ronquist R (2001) MrBayes: Bayesian inference of phylogeny. Bioinformatics 17: 754-755.

39. DeGiorgio M, Degnan J (2014) Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Syst Biol 63: 66–82.

40. Redelings B, Suchard M (2005) Joint Bayesian estimation of alignment and phylogeny. Syst Biol 54: 401-418.

41. Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science 324: 1561-1564.

42. Liu K, Warnow T, Holder M, Nelesen S, Yu J, et al. (2011) SATé-II: Very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Syst Biol 61: 90-106.

43. Mirarab S, Nguyen N, Warnow T (2014) PASTA: ultra-large multiple sequence alignment. In: Proc. Research in Computational Molecular Biology (RECOMB). pp. 177-191.

44. Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, et al. (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences 111: E4859-E4868.

45. Chifman J, Kubatko L (2014) Quartet Inference from SNP Data Under the Coalescent Model. Bioinformatics : btu530.

46. Dasarathy G, Nowak R, Roch S (2014). Data requirement for phylogenetic inference from multiple loci: a new distance method. Arxiv publication arXiv:1404.7055.

47. Dasarathy G, Nowak R, Roch S (2014) New sample complexity bounds for phylogenetic inference from multiple loci. In: IEEE International Symposium on Information Theory (ISIT). pp. 2307-2041.

48. Rodriguez F, Oliver J, Marin A, Medina J (1990) The general stochastic model of nucleotide substitution. Journal of Theoretical Biology 142: 485-501.

49. Kullback S, Leibler RA (1951) On information and sufficiency. The Annals of Mathematical Statistics 22: 79–86.

50. Fuglede B, Topsoe F (2004) Jensen-Shannon divergence and Hilbert space embedding. In: IEEE International Symposium on Information Theory. p. 31. doi:10.1109/ISIT.2004.1365067.

51. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.

52. Sukumaran J, Holder MT (2010) Dendropy: a Python library for phylogenetic computing. Bioinformatics 26: 1569-1571.

Supplementary Figures

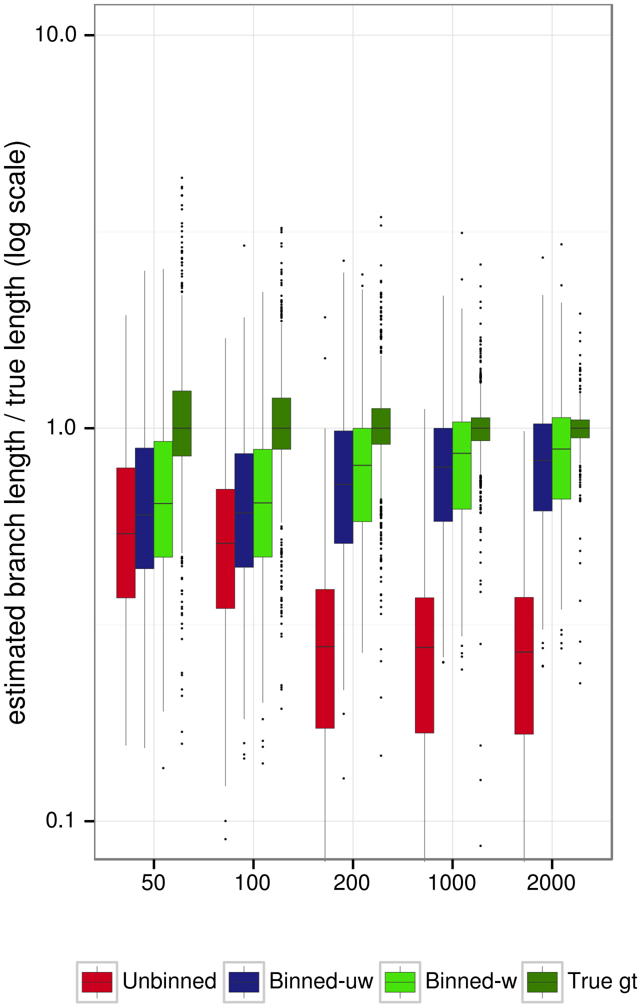


Figure S1. Effect of binning on the branch lengths (in coalescent units) estimated by MP-EST using MLBS on the avian simulated datasets with varying numbers of gene trees. We show the species tree branch length error (the ratio of estimated branch length to true branch length for branches of the true tree that appear in the estimated tree; 1 indicates correct estimation). We varied the number of genes from 50 to 2000, and fixed the sequence length to 500bp with moderate amount of ILS (1X level).

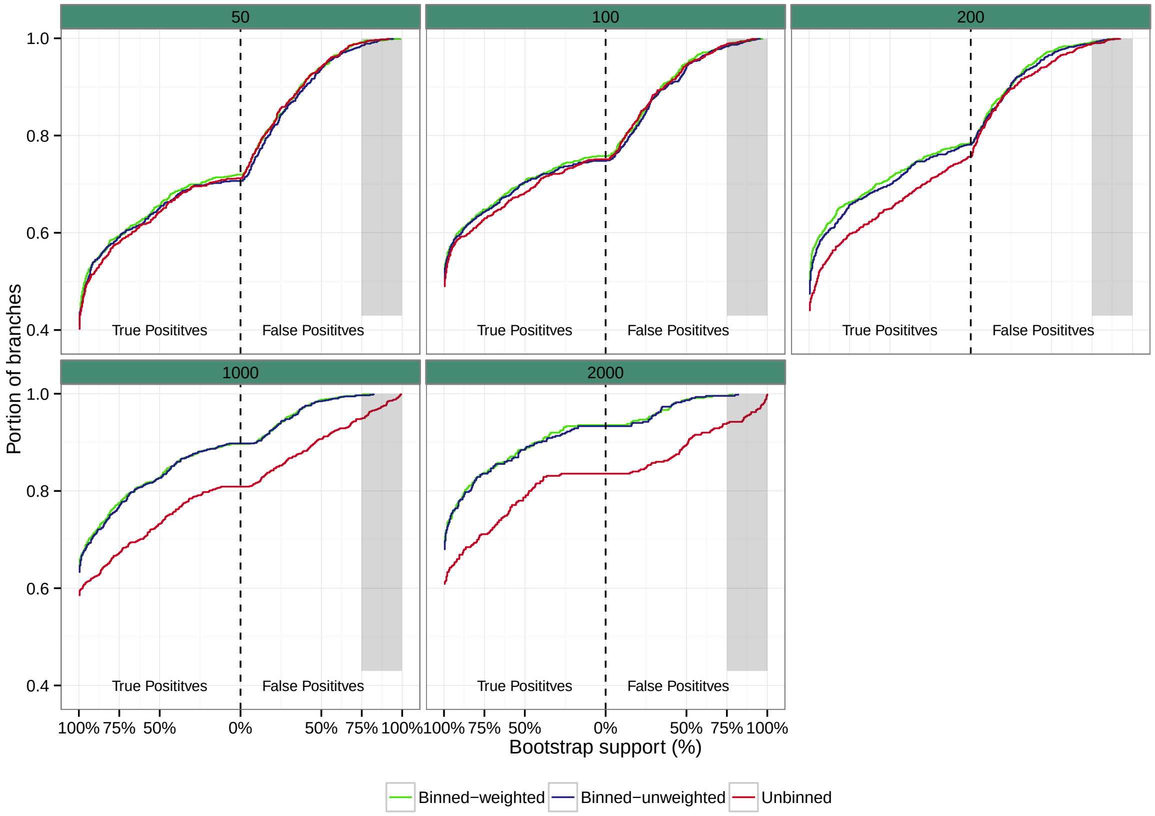
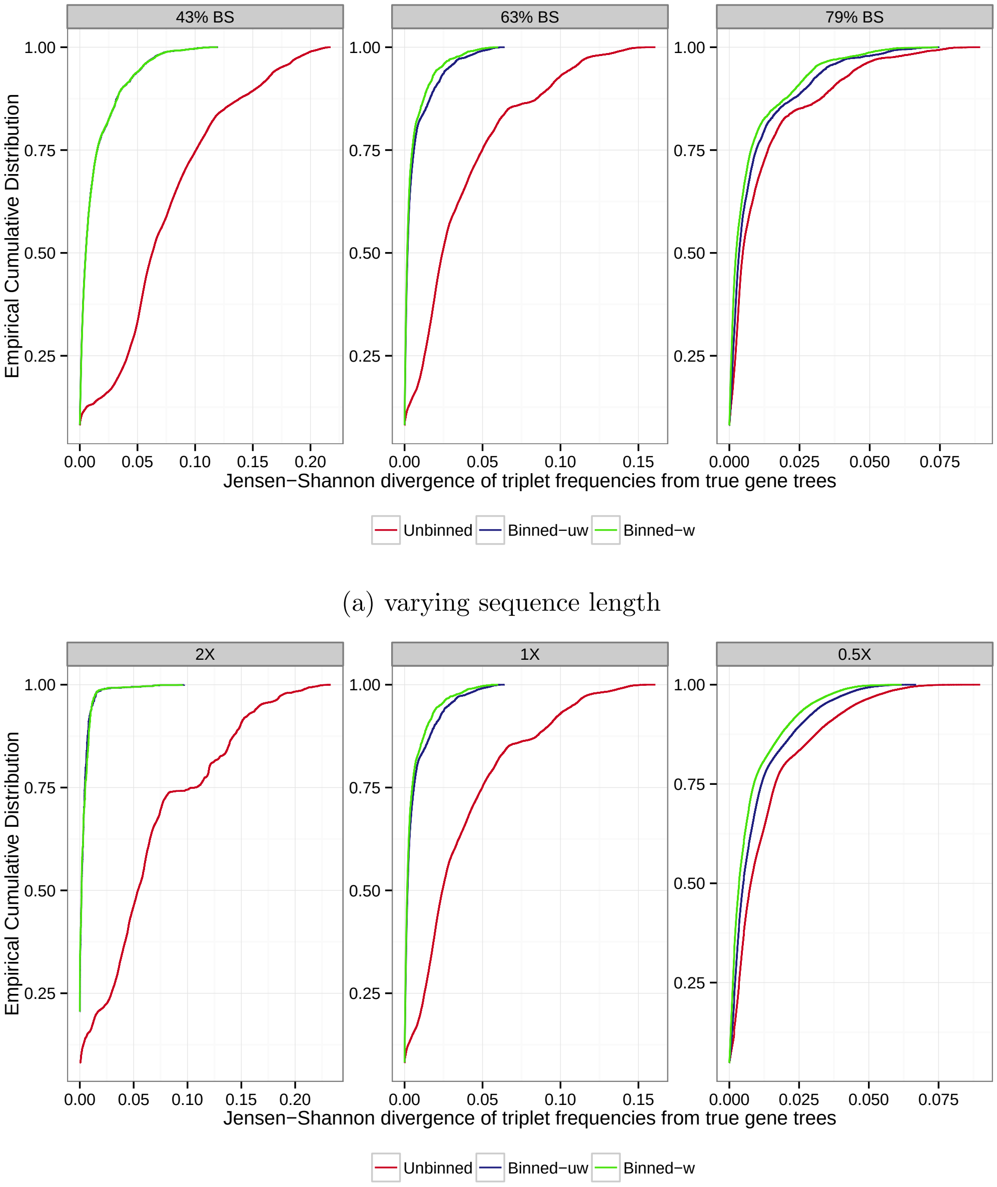


Figure S2. Cumulative distribution of the bootstrap support values (obtained using MLBS) of true positive (TP) and false positive (FP) edges estimated by MP-EST on avian datasets. We varied the numbers of genes, and fixed the sequence length to 500bp (UCE-like) with moderate amount of ILS (1X level). To produce the graph, we order the branches in the estimated species tree by their quality, so that the true positives with high support come first, followed by lower support true positives, then by false positives with low support, and finally by false positives with high support. The false positive branches with support above 75% are the most troublesome, and that fraction are indicated in the grey area. When the curve for a method lies above the curve for another method, then the first method has better bootstrap support.



(a) varying sequence length

(b) varying ILS

Figure S3. Divergence of estimated gene trees triplet distributions from true gene tree distributions for simulated mammalian datasets. (a) varying gene sequence alignments lengths with 200 number of genes and moderate levels of ILS (1X); (b) varying ILS levels with fixed 200 genes and sequence length fixed to 500bp (63% BS). 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 from estimated gene/supergene trees. For each of these $\binom{n}{3}$ triplets, we calculate the Jensen-Shannon divergence of the estimated triplet distribution from the true gene tree triplet distribution. We show the empirical cumulative distribution of these divergences. The empirical cumulative distribution shows that for a given divergence level, what percentage of the triplets are diverged from true triplet distribution at or below that level. Results are shown for 10 replicates.

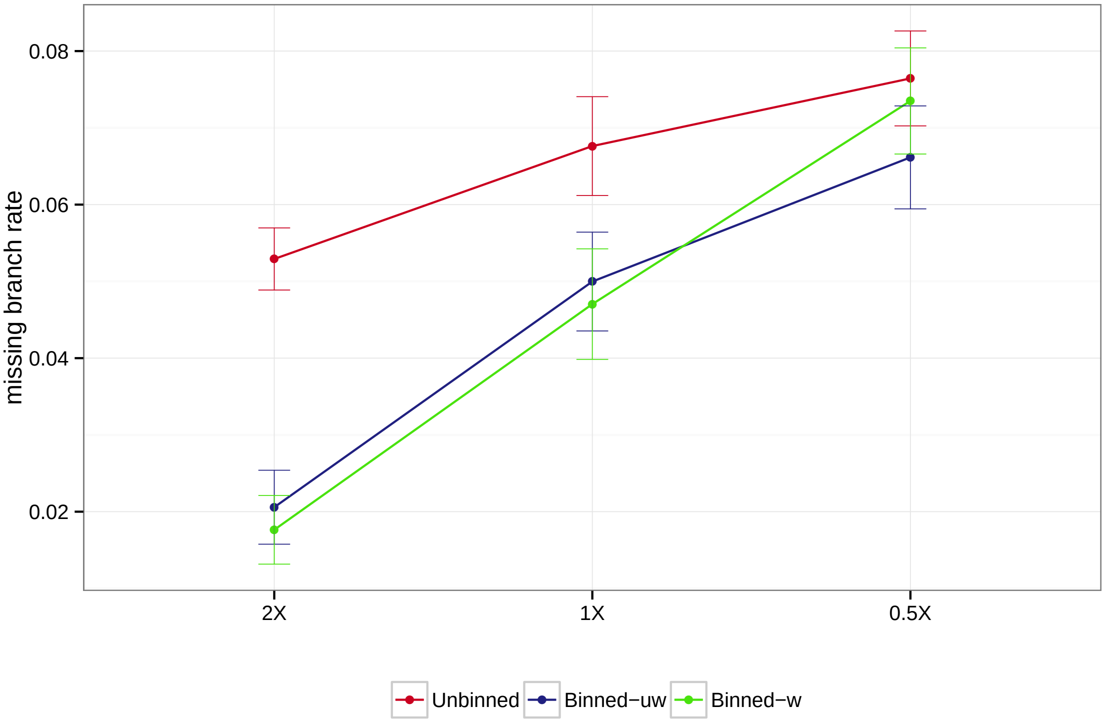


Figure S4. Species tree estimation error for MP-EST with MLBS on mammalian simulated datasets with varying amounts of ILS. We show average FN rate over 20 replicates. We varied the amount of ILS, and fixed the number of genes to 200 and sequence length to 500bp (63% BS).

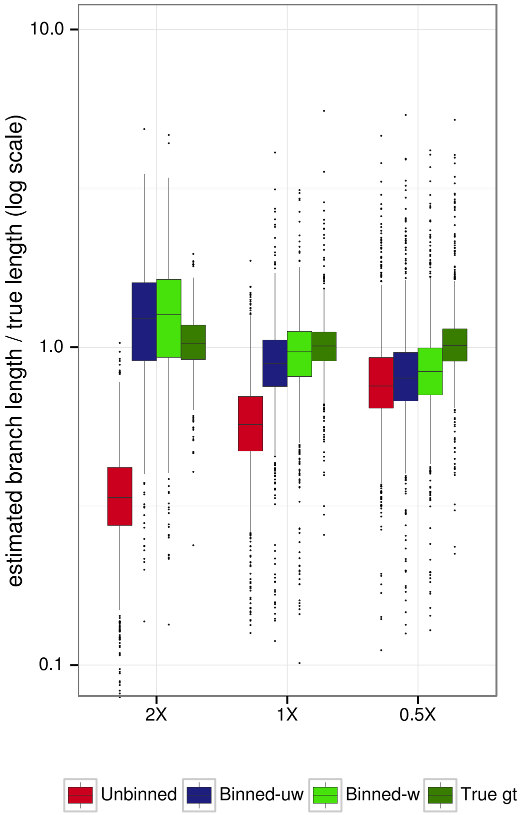


Figure S5. Effect of binning on the branch lengths (in coalescent unit) estimated by MP-EST using MLBS on the mammalian simulated datasets with varying amounts of ILS. We show the species tree branch length error (the ratio of estimated branch length to true branch length for branches of the true tree that appear in the estimated tree; 1 indicates correct estimation). We varied the amount of ILS, and fixed the number of genes to 200 and sequence length to 500bp (63% BS).

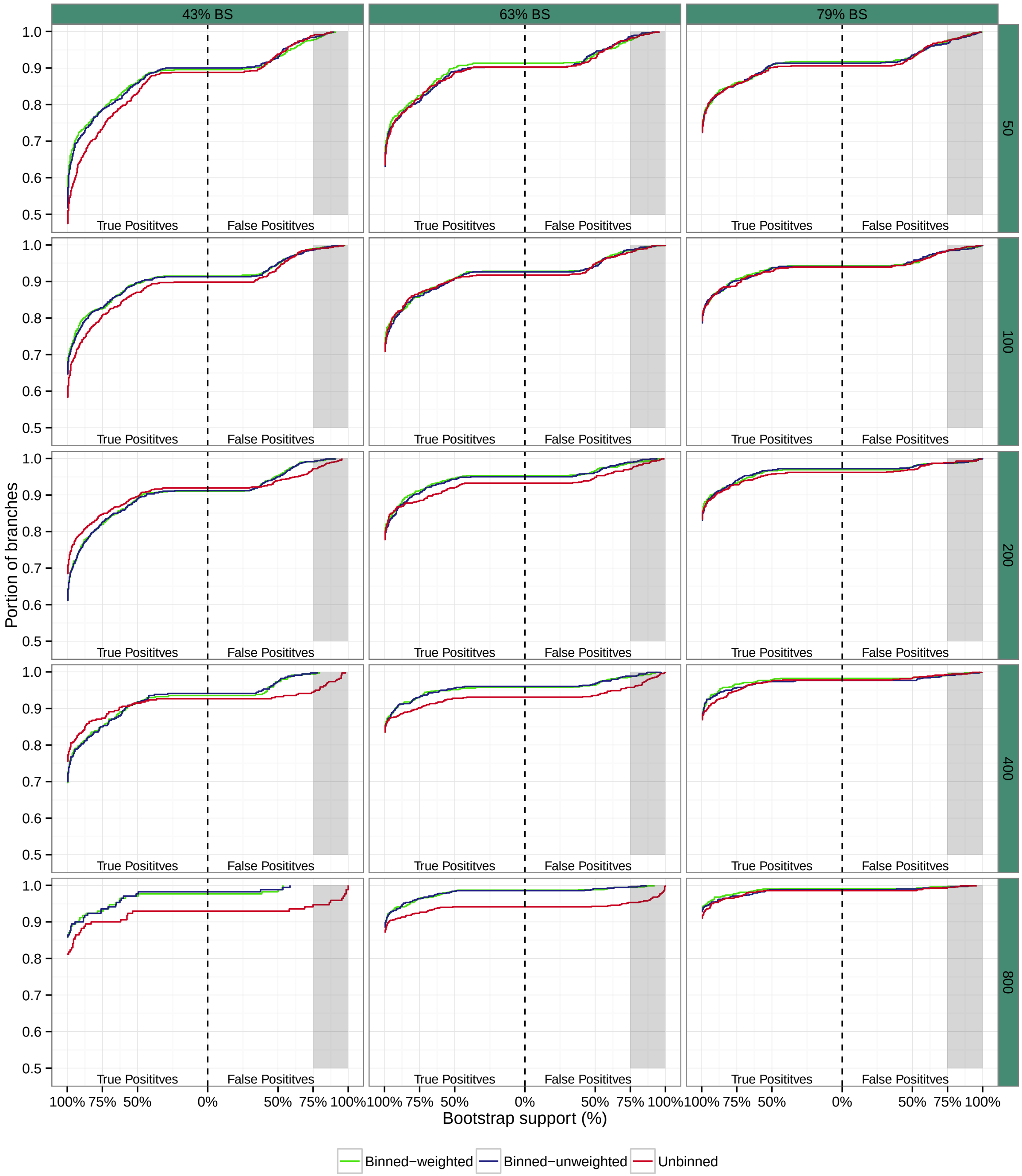


Figure S6. Cumulative distribution of the bootstrap support values (obtained using MLBS) of true positive (TP) and false positive (FP) edges estimated by MP-EST on mammalian datasets. We varied the numbers of genes, and gene sequence alignments length with moderate amount of ILS. To produce the graph, we order the branches in the estimated species tree by their quality, so that the true positives with high support come first, followed by lower support true positives, then by false positives with low support, and finally by false positives with high support. When the curve for a method lies above the curve for another method, then the first method has better bootstrap support.

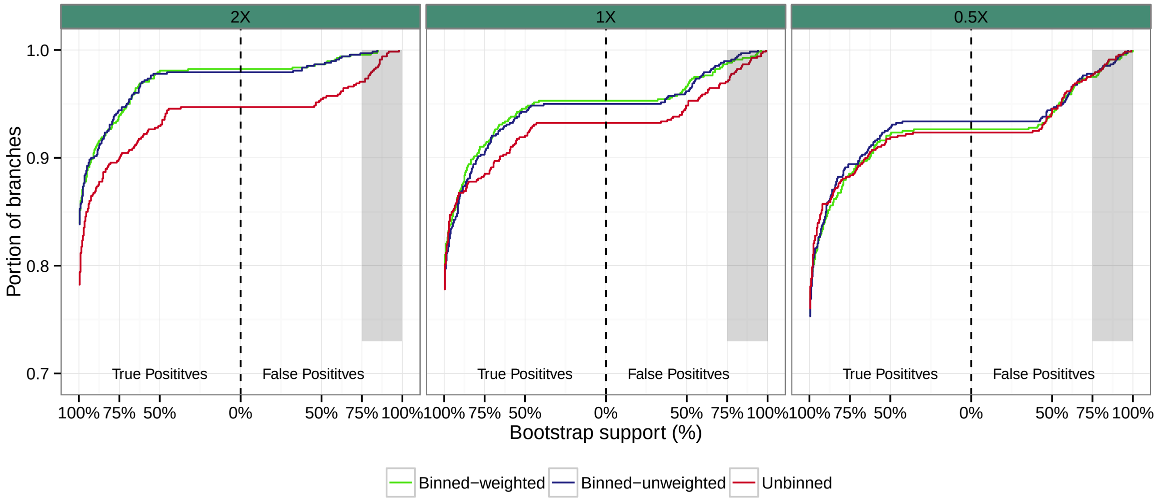


Figure S7. Cumulative distribution of the bootstrap support values (obtained using MLBS) of true positive (TP) and false positive (FP) edges estimated by MP-EST on mammalian datasets with varying amounts of ILS. We varied the amount of ILS, and fixed the number of genes to 200 and sequence length to 500bp. To produce the graph, we order the branches in the estimated species tree by their quality, so that the true positives with high support come first, followed by lower support true positives, then by false positives with low support, and finally by false positives with high support. When the curve for a method lies above the curve for another method, then the first method has better bootstrap support.