

IMPROVING GENE TREES WITHOUT MORE DATA

BY

ASHU GUPTA

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Adviser:

Professor Tandy Warnow

ABSTRACT

Species tree and gene tree estimation from sequence data are two steps in many biological analyses. Computational challenges and limited amount of data often make estimating highly accurate phylogenetic trees a difficult task. Moreover, gene alignments used to estimate trees on individual loci often have low phylogenetic signal (e.g., short alignment length), resulting in poorly estimated gene trees. Species tree estimation on the other hand is challenged by individual loci having different evolutionary histories caused by a biological phenomenon known as incomplete lineage sorting (ILS). In the presence of ILS, summary methods like MP-EST, ASTRAL2, and ASTRID are often used to estimate the species tree from gene trees. Summary methods operate by combining estimated gene trees and thus suffer in the presence of low phylogenetic signal. To tackle this problem the Statistical Binning and Weighted Statistical Binning pipelines were designed to improve gene tree estimation, which in turn can improve species tree estimation. Experimental studies of these pipelines revealed that they helped in improving gene tree and species tree estimation. However, these studies only tested the weighted statistical binning and statistical binning pipelines using multi-locus bootstrapping (MLBS) and not using BestML, where MLBS and BestML are different ways to run a phylogenetic pipeline. In this thesis, a novel phylogenetic pipeline named WSB+WQMC is proposed. This pipeline shares several design features with the weighted statistical binning pipeline (referred as WSB+CAML in this thesis) but has some other desirable properties. The WSB+WQMC pipeline is also shown to be statistically consistent under the GTR+MSC model when a slightly different version of WQMC is used.

In this study WSB+WQMC was evaluated and compared with the WSB+CAML pipeline on various simulated datasets using BestML analysis. Most of the trends seen in MLBS analyses were also observed for WSB+WQMC and WSB+CAML in BestML analyses with some important differences. It is shown that WSB+WQMC substantially improved the accuracy of gene tree and species tree estimation using ASTRAL2 and ASTRID on most datasets having low, medium, and moderately high levels of ILS. Compared to WSB+CAML, it was found that WSB+WQMC computed less accurate gene trees and species trees in certain model conditions having low and medium levels of ILS. However, WSB+WQMC was found to be better and at least as accurate as WSB+CAML in computing gene trees and species trees on all datasets having moderately high and high ILS levels. WSB+WQMC is also shown to be better in estimating gene trees on certain medium and low ILS datasets. Thus, WSB+WQMC is a potential alternative to WSB+CAML for gene tree and species tree estimation in the presence of low phylogenetic signal.

ACKNOWLEDGMENTS

I would like to express profound gratitude to my advisor Dr. Tandy Warnow for giving me the opportunity to be a part of this project and all the guidance and support she has given me along the way. I am inspired by the high standards that she sets for herself and those around her, her hard work, her attention to detail, her dedication to the profession, and her intense commitment to her work. I have thoroughly enjoyed working with her and feel like I have learned a lot.

I would like to thank my fellow lab mates: Nam Nguyen, Mike Nute, Pranjal Vachaspati, Erin Molloy, Jed Chou, and Ruth Davidson, for the stimulating discussions and for all the fun we have had in the last one year. I would also like to thank Dr. Siavash Mirarab and Md. Shamsuzzoha Bayzid for their help in my experiments.

Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

TABLE OF CONTENTS

Chapter 1: Introduction	1
1.1 Motivation	1
1.2 Background	3
1.3 WSB+WQMC	7
1.4 Figures and Tables	12
Chapter 2: Experimental Study	15
2.1 Datasets	15
2.2 Design	18
2.3 Methods	22
2.4 Figures and Tables	24
Chapter 3: Methods and Commands	26
3.1 Data availability	26
3.2 Commands	26
Chapter 4: Results	28
4.1 Training	28
4.2 Testing	29
4.3 Figures and Tables	34
Chapter 5: Discussion	56
5.1 Summary of observations	56
5.2 Impact of ILS level on WSB+WQMC and WSB+CAML	57
5.3 Impact of alignment length on WSB+WQMC	57
5.4 Impact of binning threshold on WSB+WQMC	58
5.5 Relationship between gene tree and species tree estimation error	58
5.6 MLBS vs. BestML	59
5.7 Using FastTree support values instead of bootstrap support values	59
5.8 Limitations of WSB+WQMC	59
Chapter 6: Summary	61
Bibliography	64

Chapter 1

Introduction

1.1 Motivation

Phylogenetic trees, whether of individual loci (so called “gene trees”) or genome-level (species trees) provide a basis for understanding how life evolved on earth and have many biological and medical applications. Thus, the estimation of gene trees and species trees is a basic step in many biological analyses [1]. However, computing these trees with high accuracy is very challenging due to a variety of reasons including computational challenges (almost all problems are NP-Hard) and limited amounts of data. Due to limited amount of data for individual loci, the gene alignments used in phylogenetic studies often have low phylogenetic signal (i.e., its sequences are too short or it evolves too slowly) [2–4]. Gene tree estimation together with low phylogenetic signal is also plagued by errors in gene alignment itself. Moreover, many biological conditions (e.g., short branches, long branches, etc.) also make it impossible to estimate highly accurate gene trees.

Species tree estimation on the other hand faces the problem of individual loci having different evolutionary histories, caused by the biological phenomenon known as incomplete lineage sorting (ILS) [5–8]. Furthermore, in the presence of ILS, standard methods like concatenation and consensus can be statistically inconsistent [9, 10] and are known to perform poorly under some conditions [11]. Therefore, many new methods have been developed for estimation of species trees in the presence of ILS (e.g., co-estimation methods, summary methods, etc.). Co-estimation methods, which co-estimate gene trees and species tree like BEST [12] and *BEAST [13], generally have excellent accuracy but are limited by their computational complexity. Summary methods on the other hand operate by combining estimated gene trees into a species tree. Some of the summary methods like ASTRAL2 [14], ASTRID [15] and MP-EST [16] are based on the multi-species coalescent model [17] and are statistically consistent under that model. Summary methods that are statistically consistent under the multi-species coalescent model reconstruct the true species tree with high probability given a sufficiently large number of true gene trees. Summary methods are the most

frequently used methods for estimating species trees under the multi-species coalescent model, and have shown good results on many biological datasets [12, 18].

Summary methods assume perfectly accurate gene trees for statistical consistency and thus suffer when the gene trees have low accuracy [19, 20]. Thus, an attempt to improve gene trees by enhancing phylogenetic signal not only helps in getting better gene evolutionary histories, but it also indirectly affects the accuracy of species tree computed by summary methods. One obvious solution to the problem is collecting more data, which is both costly and time consuming, making it infeasible for most studies. If collecting more data is not an option, another obvious direction to boost the phylogenetic signal of an individual gene is by using the data from other genes. Co-estimation methods like BEST and *BEAST use this idea to enhance gene tree and species tree estimation. However, their huge computational complexity limits their use in most studies.

The Naive Binning pipeline [2] was another direction for boosting phylogenetic signal and showed improvements in the accuracy of estimated species trees. In this pipeline genes are randomly grouped into bins of same size and a supergene alignment is computed for each bin by concatenating gene alignments present in that bin. The naive binning pipeline then uses maximum likelihood to get a supergene tree on each bin. These supergene trees can then be used to compute a species tree using a summary method. Statistical Binning [3] and Weighted Statistical Binning [4] improved the naive binning pipeline and grouped genes in a smarter way to ensure genes within a bin have similar evolutionary histories and less discord. Both the statistical binning and the weighted statistical binning pipelines were evaluated on various datasets and showed positive results. Weighted statistical binning was also proved to be statistically consistent under the GTR+MSC model, a model where gene trees evolve within a species tree under the multi-species coalescent (MSC) model [17] and sequences evolve down each gene tree under the General Time Reversible (GTR) model [21]. However, these pipelines were only tested using multi-locus bootstrapping (MLBS) analysis of datasets and not using BestML analysis, which is known to be more accurate than MLBS in certain model conditions having large enough number of genes [22]. Second, statistical binning and weighted statistical binning used concatenation within each bin, which is not only computationally expensive but may also lead to poor gene tree accuracy if gene trees within a bin have different evolutionary histories. Third, statistical binning and weighted statistical binning discarded the topology of the initially estimated gene tree, which potentially can be fairly accurate and close to the true evolutionary history in many cases. Finally, statistical binning and weighted statistical binning computed a single supergene tree for each bin rather than a unique gene tree for each input gene tree.

In this thesis a novel phylogenetic pipeline named WSB+WQMC is proposed. It aims to improve gene tree and species tree estimation when the individual loci have low phylogenetic signal. It modifies the weighted statistical binning pipeline and computes a new unique gene tree for each initial gene tree. The new pipeline

uses WQMC [23] (a quartet-based tree estimation method) to compute a new gene tree from a set of weighted quartets unique to each gene. In the WSB+WQMC pipeline, quartets from initially estimated gene trees (i.e., original gene trees) within a bin are combined uniquely for each gene by up-weighting its own quartets; hence, it allows WSB+WQMC to use the topology of initially estimated gene tree to its advantage. The WSB+WQMC pipeline has good theoretical properties and is shown to be statistically consistent under the GTR+MSC model when a slightly different version of WQMC is used in the pipeline. The WSB+WQMC pipeline is tested on BestML analysis of various datasets and the results are compared with the original weighted statistical binning pipeline (which we refer to as WSB+CAML) proposed in [4].

1.2 Background

Basic phylogenetic pipeline using a summary method (Unbinned)

The basic unbinned phylogenetic pipeline using a summary method is shown in Figure 1.1. The pipeline begins with a set of sequences for different loci across different species as input. Gene alignments for each gene are then computed using any alignment method. Next, gene trees are estimated on each locus using the input gene alignments. Finally, it combines the estimated gene trees into a species tree using a summary method.

MLBS vs. BestML

Multi-locus bootstrapping (MLBS) and BestML are different ways to run a phylogenetic pipeline that uses a summary method. When running the pipeline using BestML analysis, a single gene tree having the best maximum likelihood score is estimated for each gene. Then, the best maximum likelihood gene tree estimate for each gene is used by a summary method to compute a species tree.

Given n genes having k bootstrap replicated gene alignments each, the phylogenetic pipeline using MLBS analysis is run using the following steps.

- For each gene i and bootstrap replicated alignment j , a tree $t_{i,j}$ is estimated using a maximum likelihood method with j^{th} bootstrap replicated gene alignment for gene i as input. Here $1 \leq i \leq n$ and $1 \leq j \leq k$.
- Then, for $j = 1$ up to k , a summary method is run with the j^{th} bootstrap replicated alignment for each gene as input (i.e., $\{t_{1,j}, t_{2,j}, \dots, t_{n,j}\}$ is the input to the summary method in j^{th} run) to compute a species tree t_j^{sp} .

- Finally, a greedy consensus tree of k species trees from the previous step (i.e., $\{t_1^{sp}, t_2^{sp}, \dots, t_k^{sp}\}$) is computed. This greedy consensus tree is the output species tree.

Multi-locus bootstrapping (MLBS) is more frequently used in phylogenetic analyses. However, MLBS is not necessarily the best way to compute gene trees and species trees. In [22], the accuracy of species tree estimation methods like Greedy consensus, MRP [24], MRL [25], and MP-EST [16] using MLBS analysis was compared to the accuracy of species tree estimation of these methods using BestML analysis. It was found that all methods computed more accurate species tree using MLBS analysis when the number of genes used was small. On the other hand, when a large enough number of genes were used, it was found that all methods computed more accurate species tree using BestML analysis. This highlights the importance of testing and evaluating all phylogenetic pipelines using both MLBS and BestML analyses.

Naive Binning

The benefit of using binning-based pipelines to improve gene tree estimation has only recently been discovered. The first study that explored binning was [2], which introduced the naive binning pipeline. In this technique, genes were randomly partitioned into disjoint bins and supergene alignments for each bin were computed by concatenating individual gene alignments present in that bin. Maximum likelihood trees were computed for each supergene alignment, and then these supergene trees were combined into a species tree using a summary method.

The naive binning pipeline was evaluated using BestML analysis on simulated datasets with species tree estimation methods including *BEAST, concatenation, and several summary methods. It was observed that in low ILS conditions, the naive binning pipeline improved the accuracy of species tree estimation using summary methods and improved scalability of *BEAST without impacting its accuracy. These observations showed that naive binning can help in improving species tree estimation when the individual genes have low phylogenetic signal. However, since naive binning puts genes into bins randomly and can lead to genes with very different evolutionary histories put in the same bin, it was conjectured that it could reduce accuracy in high levels of ILS.

Statistical Binning

Statistical Binning [3] improved on the naive binning pipeline and used a different binning technique than random binning. The new binning technique used bootstrap support values on the estimated gene trees to partition the set of loci into bins. The binning technique created bins of approximately equal sizes and ensured that no two genes within the same bin have conflicting edges having high support values in their

respective tree topologies. This smarter binning technique ensured that genes with different evolutionary histories were unlikely to be placed in the same bin. Similar to naive binning, supergene alignments were computed for each bin by concatenating gene alignments within that bin. Then, supergene trees computed on the supergene alignments using a maximum likelihood gene tree estimation method were used to compute a species tree using a summary method.

The statistical binning pipeline was evaluated in [3] on many biological and simulated datasets using MLBS analysis and MP-EST as the summary method. It was observed that this pipeline helped in reducing error in the MP-EST species tree topology on all the datasets studied. It was also observed that the relative improvement in the accuracy of MP-EST species tree with respect to the basic unbinned pipeline was more when the level of ILS was low. However, the statistical binning pipeline was only tested on datasets having 37 or more taxa and using only MP-EST as the summary method. Moreover, statistical binning wasn't evaluated on datasets having ILS levels more than 60% AD. In [3], the theoretical properties of statistical binning were also not addressed. Subsequently, Theorem 3 from [4] showed that the statistical binning pipeline is statistically inconsistent under the GTR+MSC model.

Weighted Statistical Binning (WSB+CAML)

The weighted statistical binning pipeline [4] (referred as WSB+CAML in this thesis) modified the statistical binning pipeline by adding one more step before running a summary method on supergene trees. It repeated the supergene trees as many times as the number of genes in its bin, and used those repeated supergene trees as the input for a summary method to compute a species tree. Detailed steps in the weighted statistical binning pipeline are shown in Figure 1.2. The input to the WSB+CAML pipeline is a set of gene alignments on different loci. The WSB+CAML pipeline then proceeds with the following steps.

- Step 1: It computes a maximum likelihood (ML) gene tree with bootstrap support values for each gene from its sequence alignment. These gene trees are also referred as initial or original gene trees.
- Step 2 (Statistical Binning): It computes an incompatibility graph based on a bootstrap threshold t , with each gene corresponding to a vertex in the graph. For each pair of vertices in the graph, an edge is added between them if the gene trees corresponding to the vertices are incompatible with each other. To determine incompatibility between gene trees, first all the edges having bootstrap support values less than t are collapsed, and then the compatibility is checked between the edge-collapsed gene trees. Two gene trees are said to be compatible with each other if they have a common refinement. After that, a heuristic from [3] is used to color the vertices of the incompatibility graph such that no two adjacent nodes have the same color, the number of distinct colors used is small, and the color classes

are approximately of the same size (minimum balanced vertex coloring problem). The coloring of the nodes defines the output bins with genes having same color in the same bin.

- Step 3: For each bin, gene alignments of its constituent genes are concatenated and define the supergene alignment.
- Step 4: A supergene tree is then estimated for each supergene alignment using a fully partitioned maximum likelihood gene tree estimation method.
- Step 5: Each supergene tree is repeated as many times as the number of genes in its bin. For each gene, the supergene tree for its bin is considered as its new gene tree.
- Step 6: The new gene trees computed in step 5 are used as input to estimate a species tree using a summary method.

The WSB+CAML pipeline was found to have strong theoretical guarantees. Theorems 1 and 2 from [4] showed that this small change in the statistical binning pipeline of repeating supergene trees made the new weighted statistical binning pipeline statistically consistent under the GTR+MSC model.

Theorem 1 from [4]: Let T^{sp} be a species tree with branch lengths in coalescent units, and $\mathcal{T} = \{t_1, t_2, \dots, t_p\}$ be a set of p rooted gene trees sampled from the distribution defined by T^{sp} under the multi-species coalescent model. Let $\{\theta_1, \theta_2, \dots, \theta_p\}$ be a set of numeric GTR model parameters (gene tree branch lengths and 4×4 substitution matrices) so that $T_i = (t_i, \theta_i)$ is a GTR model tree for each $i = 1, 2, \dots, p$. Let $\mathcal{T}' = \{T_1, T_2, \dots, T_p\}$. For each i , $1 \leq i \leq p$, let sequence dataset S_i evolve down the GTR model tree T_i . Let $\epsilon < 1$ and bootstrap support threshold $B < 1$ be given. Then, there is a sequence length L (that depends on $\mathcal{T}'$ and ϵ) such that if at least L sites evolve down each gene tree, then with probability at least $1 - \epsilon$, the following will be true.

- For each $i = 1, 2, \dots, p$, the gene tree estimated using a GTR maximum likelihood method on S_i will have the same unrooted topology as t_i (the true gene tree for S_i) and will have bootstrap support values greater than B for all its branches.
- For every bin produced by the binning technique used in WSB+CAML based on GTR maximum likelihood analyses of the gene sequence alignments, the estimated gene trees for genes in the bin will have the same topology, and
- All genes with the same true gene tree topology will be in the same bin.

This theorem is copied directly from [4] and presented here as it is later used in the statistical consistency proof of WSB+WQMC.

Theorem 2 from [4]: The phylogenetic pipeline that uses GTR maximum likelihood to estimate gene trees, uses weighted statistical binning to compute supergene trees, and then combines the supergene trees using a coalescent-based summary method, is statistically consistent under the GTR+MSC model.

Note that the term “coalescent-based summary method” in this theorem refers to summary methods that are statistically consistent under the multi-species coalescent model.

In [4], the weighted statistical binning pipeline was evaluated and compared with the statistical binning pipeline on various biological and simulated datasets using MLBS analysis and two summary methods, ASTRAL2 and MP-EST. It was found that statistical binning and weighted statistical binning had no significant differences in terms of species tree estimation on the datasets studied. It was also found that weighted statistical binning helps in improving species tree topologies except for a small number of datasets having very high levels of ILS and small numbers of taxa. The relative magnitude of improvement in the accuracy of species trees compared to the basic unbinned pipeline was also seen to be more for conditions having low ILS. The difference between species tree estimation error by running the WSB+CAML pipeline and the basic unbinned pipeline decreased with increasing sequence length.

1.3 WSB+WQMC

In this section, a novel phylogenetic pipeline named WSB+WQMC is presented, which aims to improve the accuracy of gene tree and species tree estimation compared to the weighted statistical binning pipeline. This pipeline modifies the WSB+CAML pipeline and avoids computing supergene alignments and supergene trees. Rather, it computes a set of weighted quartets for each gene using quartets induced by its original gene tree and other quartet topologies from genes within its own bin. It computes a unique weighted quartet set for a gene by combining the weighted quartets induced by other original gene trees within its bin and weighted quartets induced by itself (up-weighted by a factor of confidence value $c \times \text{size of bin}$). These weighted quartet sets are then used to compute the new gene trees using WQMC [23]. These new gene trees are then used to estimate a species tree by a summary method. Detailed steps in the WSB+WQMC pipeline are shown in Figure 1.3. The input to the WSB+WQMC pipeline is a set of gene alignments on different loci. The WSB+WQMC pipeline then proceeds with the following steps.

- Step 1: Same as step 1 in the WSB+CAML pipeline.

- Step 2 (Statistical Binning): Same as step 2 in the WSB+CAML pipeline.
- Step 3: Given a confidence value c (algorithmic parameter), for each gene, a set of weighted quartets is computed. Let $\mathcal{G}$ be the set of original gene trees and $B_1, B_2, \dots, B_k$ be the bins computed in step 2, where k is the number of bins.

Consider a gene $g_1 \in \mathcal{G}$, WLOG $\{g_1, g_2, \dots, g_{|B_l|}\} = B_l$.

- First the set $\mathcal{Q}_{g_1} = \{\langle q, upweight * w_{g_1}(q) \rangle \mid q \text{ is a quartet topology induced by gene tree } g_1\}$ is computed. Here $w_g(q) = \frac{\text{length of internal branch of quartet } q \text{ induced by } g}{\text{diameter of the quartet } q}$, and $upweight = c * |B_l|$, where confidence value c is an algorithmic parameter. Note that the maximum likelihood tree comes with branch lengths that are non-negative and hence leaf-to-leaf distances are defined by the branch lengths. The diameter of a quartet is the largest distance between any two leaves in the tree. The set $\mathcal{Q}_{g_1}$ has a tuple (quartet topology and its weight) for all the quartet topologies the gene tree g_1 induces.
- Second, for each other gene tree in the bin, $t \in \{B_l \setminus g_1\}$, the set $\mathcal{Q}_t = \{\langle q, w_t(q) \rangle \mid q \text{ is a quartet topology induced by the gene tree } t\}$ is computed.
- Finally, all the sets $\mathcal{Q}_{g_1}, \mathcal{Q}_{g_2}, \dots, \mathcal{Q}_{g_{|B_l|}}$ are merged into a single weighted quartet set $\mathcal{M}_{g_1}$, where the weight of any quartet topology is the sum of weights of the same quartet topology in all other sets.

These steps are repeated to obtain all weighted quartet sets $\mathcal{M}_{g_1}, \mathcal{M}_{g_2}, \dots, \mathcal{M}_{|\mathcal{G}|}$.

The weight function w used to determine the weight of the quartet was used in [23] and is a measure of reliability of the quartet.

- Step 4: The new gene tree is computed for each gene by running WQMC on its weighted quartet set as input.
- Step 5: The new gene trees computed in step 4 are used as input to estimate a species tree using a summary method.

The basic intuition behind WSB+WQMC is to use the frequent quartet topologies present in the bin to possibly rectify incorrect quartet topologies induced by an original gene tree. The confidence value c used in the pipeline determines the robustness to change from the original gene tree topology. The larger the confidence value c , the more frequent a conflicting quartet topology must occur in the bin to change the

gene tree topology. Thus, by having this control over robustness to change, a user can control the extent of deviation of new gene tree from the original gene tree. The binning threshold value t used in step 2 on the other hand affects the size of bins. A large binning threshold results in larger bin sizes with only a single bin for $t = 100\%$ (referred as one-bin), whereas a smaller binning threshold results in smaller bin sizes with singleton bins for $t = 0\%$. Determining both the optimal binning threshold t and confidence value c is a very challenging problem and affects the performance of the pipeline.

Proof of statistical consistency of WSB+WQMC when WQMC used in the pipeline is replaced by WQMC*

In this section, it is proved that WSB+WQMC is statistically consistent under the GTR+MSC model when a slightly different version of WQMC is used in the pipeline. For the purpose of this proof, WQMC used in the WSB+WQMC pipeline is replaced by WQMC*. WQMC* is a slightly different version of WQMC with an additional property that it exactly solves the trivial maximum weighted quartet compatibility (MWQC) problem (defined below). The statistical consistency proof given in this section is in the same lines as the statistical consistency proof described in [4] for WSB+CAML. The WSB+WQMC pipeline uses the same binning technique as used in WSB+CAML and therefore Theorem 1 from [4] also applies to WSB+WQMC. The main result of this section is described in Theorem 2. The trivial MWQC problem is solvable in polynomial time using the All Quartets Method (defined below) and therefore replacing WQMC with WQMC* in the WSB+WQMC pipeline for the purpose of this proof is reasonable.

Definitions

Maximum weighted quartet compatibility problem (MWQC): Given a set $Q = \{\langle q, w(q) \rangle \mid q \text{ is a quartet topology and } w \text{ is any positive weight function}\}$, find a tree t that maximizes the weight $W(t)$, where $W(t) = \sum_{x \in Q'(t)} w(x)$, where $Q(t)$ is a set containing all the quartet topologies induced by the tree t , and $Q'(t) = Q \cap Q(t)$.

Trivial weighted quartet set: A weighted quartet set $Q = \{\langle q, w(q) \rangle \mid q \text{ is a quartet topology and } w \text{ is any positive weight function}\}$ is called a trivial weighted quartet set when it contains all quartet topologies induced by a tree t and nothing else.

Trivial MWQC problem: An instance of MWQC problem is said to be trivial when the set Q is a trivial weighted quartet set from a tree t^* . It can be easily seen that the solution to the trivial MWQC

problem is the tree t^* itself

All Quartets Method (This is taken from [26])

The input to the method is a weighted quartet set Q . Let $S = \{s_1, s_2, \dots, s_n\}$ be the leaf set of the quartet trees in Q . It is assumed that $|S| \geq 4$, since otherwise there are no quartets. Also, it is assumed that the quartet set Q contains exactly one quartet tree on every four leaves. Otherwise, the All Quartet Method fails to compute a tree and returns “No compatibility tree”. Given the input set Q , the All Quartets Method proceeds with the following steps.

- if $|S| = 4$, then return the quartet tree in Q . Else, find a pair s_i, s_j that are always grouped together in any quartet that includes both s_i, s_j . If no such pair exists, return “No compatibility tree” and exit. Otherwise, remove all the quartets that include both s_i, s_j from the set Q .
- Recursively compute a tree t' on $S - \{s_i\}$.
- Return the tree created by inserting s_i next to s_j in t' .

Note: It can be easily seen that the All Quartets Method returns the tree t , when the trivial weighted quartet set from a tree t is given as input.

WQMC*: WQMC* is a slightly different version of WQMC and proceeds as follows.

- Given an input set of weighted quartets Q , the All Quartets Method is run with the set Q as input. If the All Quartets Method returns a compatibility tree t , return t and exit. Otherwise, proceed to the next step.
- Run WQMC with the set Q as input and return the output tree.

Corollary 1: Let T^{sp} be a species tree with branch lengths in coalescent units, and $\mathcal{T} = \{t_1, t_2, \dots, t_p\}$ be a set of p rooted gene trees sampled from the distribution defined by T^{sp} under the multi-species coalescent model. Let $\{\theta_1, \theta_2, \dots, \theta_p\}$ be a set of numeric GTR model parameters (gene tree branch lengths and 4×4 substitution matrices) so that $T_i = (t_i, \theta_i)$ is a GTR model tree for each $i = 1, 2, \dots, p$. Let $\mathcal{T}' = \{T_1, T_2, \dots, T_p\}$. For each i , $1 \leq i \leq p$, let sequence dataset S_i evolve down the GTR model tree T_i . Let $\epsilon < 1$ and bootstrap support threshold $B < 1$ be given. Then, there is a sequence length L (that depends on $\mathcal{T}'$ and ϵ) such that if at least L sites evolve down each gene tree, then with probability at least $1 - \epsilon$, the following will be true.

- For each $i = 1, 2, \dots, p$, the gene tree estimated using a GTR maximum likelihood method on S_i will have

the same unrooted topology as t_i (the true gene tree for S_i) and will have bootstrap support values greater than B for all its branches.

- For every bin produced by the binning technique used in WSB+WQMC based on GTR maximum likelihood analyses of the gene sequence alignments, the estimated gene trees for genes in the bin will have the same topology, and
- All genes with the same true gene tree topology will be in the same bin.

Proof: The bins produced by WSB+CAML and WSB+WQMC are identical. Therefore, we have this corollary that is identical to Theorem 1 from [4] except that WSB+CAML is replaced by WSB+WQMC.

Theorem 2: The WSB+WQMC phylogenetic pipeline with WQMC replaced by WQMC* is statistically consistent under the GTR+MSC model.

Proof: By Corollary 1, as the sequence length for each gene tree goes to infinity, all the genes put in any bin by WSB+WQMC will have the same true gene tree topology with probability converging to 1. This will cause the weighted quartet sets for each gene g_i to be a trivial weighted quartet set with quartet topologies only from the true gene tree t_i . Since it has been assumed that WQMC* used to compute a gene tree from a weighted quartet set solves the trivial MWQC exactly, the new gene trees will have the same topology as their corresponding true gene tree. Hence, the distribution of new gene trees produced using WSB+WQMC will be identical to the distribution of the true gene trees for these genes. Therefore, as number of sites and number of genes increase, the gene tree distribution from WSB+WQMC gene trees converges to the true gene tree distribution. Finally, as the gene tree distribution converges to the true gene tree distribution, the species tree computed by a statistically consistent summary method also converges to a true species tree.

1.4 Figures and Tables

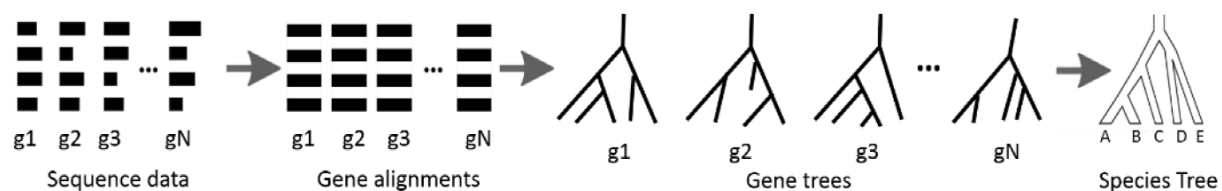


Figure 1.1: **Basic unbinned phylogenetic pipeline using a summary method.** The input to the pipeline is a set of sequences for different loci across different species. In this pipeline, a multiple sequence alignment is computed using any alignment method. Then a gene tree is computed on each gene using the multiple sequence alignment. Gene trees are then used by a summary method to compute a species tree.

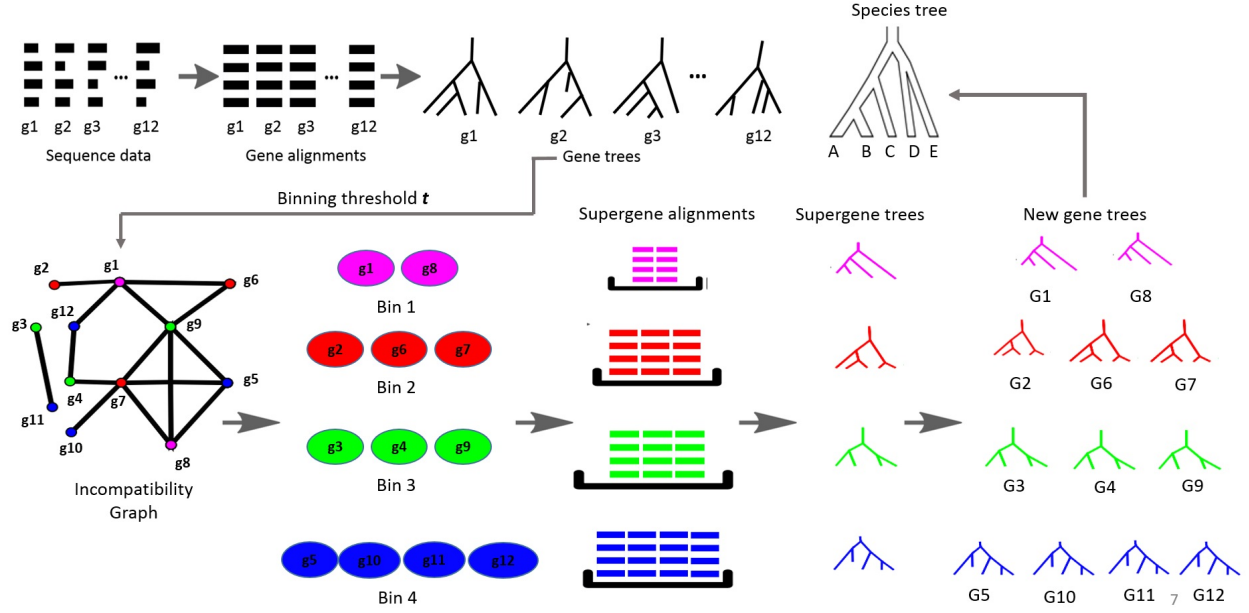


Figure 1.2: **WSB+CAML pipeline for phylogenetic analysis.** The input to the pipeline is a set of gene alignments computed from a set of sequences for different loci across different species. Gene trees are then estimated on the gene alignments using a maximum likelihood tree estimation method. Gene trees are then used to compute an incompatibility graph, where each vertex represents a gene and each edge represents incompatibility between them based on binning threshold t . A heuristic for balanced minimum vertex coloring is run to divide the genes into disjoint bins. For each bin, gene alignments for genes within that bin are concatenated into a supergene alignment. Supergene trees are then estimated using a fully partitioned maximum likelihood gene tree estimation method. Supergene tree for each bin is repeated for as many genes in that bin and considered as the new gene tree. New gene trees are used by a summary method to compute a species tree. Please note that the example used to describe the WSB+CAML pipeline is taken from Figure 1 of [4].

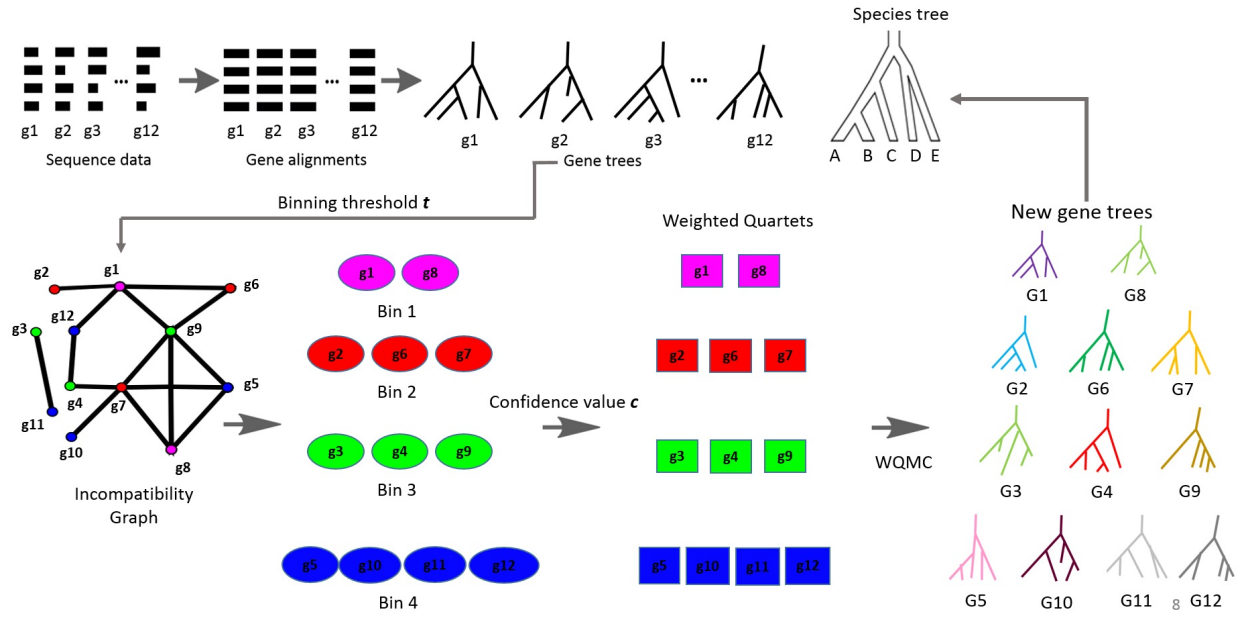


Figure 1.3: **WSB+WQMC pipeline for phylogenetic analysis.** The input to the pipeline is a set of gene alignments computed from a set of sequences for different loci across different species. Gene trees are then estimated on the gene alignments using a maximum likelihood tree estimation method. Then, genes are divided into disjoint bins using the binning technique used in WSB+CAML with binning threshold t . For each gene, a set of weighted quartets is computed by combining weighted quartet topologies within its bin with up-weighting own quartets by confidence value c . WQMC is then run with weighted quartet set as input to get a new gene tree for each gene. New gene trees are then used by a summary method to compute the species tree.

Chapter 2

Experimental Study

2.1 Datasets

In the experimental study, datasets were separated into two disjoint sets: a training dataset and a testing dataset. Table 2.1 lists the datasets used for training and Table 2.2 lists the datasets used for testing. The ILS level in the datasets was measured using AD%, which is the average percentage of missing branches between true gene trees and the true species tree. Higher value of AD% indicates higher level of ILS in the dataset. For the purpose of this study, the range of ILS levels was broken down into four different categories, namely low, medium, moderately high, and high, as shown in Table 2.3. Note that the high level of ILS has a very big range and there are probably huge differences between the low end of the high ILS range and the high end of the high ILS range.

The mammalian simulated dataset studied in [3] (based on an MP-EST [16] analysis of a mammalian biological dataset having 37 taxa) was used for the training phase. For the testing phase, collections of simulated datasets with 10, 11, 15, and 50 taxa along with the Mammalian-1X dataset having 250bp alignment length were used. The ILS level in the simulated datasets used in the study ranged from low ILS (11-taxon L ILS and SimPhy-10M10e7) to high ILS conditions (10-taxon H ILS, 15-taxon H ILS, SimPhy-500K10e7, and SimPhy-500K10e6). All the datasets except the SimPhy datasets had bootstrap support values on the original gene trees available from earlier studies. The 15-taxon datasets evolve under a strict molecular clock, but the other datasets do not.

The original gene trees for the Mammalian, 10-taxon, 15-taxon and 11-taxon datasets were computed in prior studies [2–4] from the gene alignments using maximum likelihood estimation method RAxML version 7.3.5 [27]. To compute the original gene trees from gene alignments, the maximum likelihood method FastTree 2.1 [28] was used for the SimPhy datasets. These maximum likelihood methods were also used to estimate supergene trees from the supergene alignments in the WSB+CAML pipeline. RAxML version 8.2.7 was used in unpartitioned mode for computing supergene trees from the supergene alignments for the

mammalian dataset. For the 10-taxon, 11-taxon and 15-taxon datasets, RAxML version 8.2.7 was used in fully partitioned mode to compute supergene trees from the supergene alignments. To compute supergene trees from the supergene alignments for the SimPhy dataset, FastTree 2.1 was used in unpartitioned mode. More detailed information about each dataset is presented in the following paragraphs.

2.1.1 Mammalian

This dataset was also studied in [3]. This dataset is based on an MP-EST analysis of a mammalian biological dataset and has 37 taxa. To vary the amount of ILS, the branch lengths in the species tree estimated using MP-EST on biological data were modified. To simulate higher level of ILS the branch lengths were shortened by dividing by 2 (Mammalian-0.5X) and to simulate lower level of ILS branch lengths were multiplied by 2 (Mammalian-2X). For the Mammalian-1X datasets, the branch lengths were kept unchanged. This datasets ranges in ILS level from low (21% AD) for the 2X branch length to high (50% AD) for 0.5X branch length. The Mammalian-1X dataset provides multiple sequence alignment of lengths 250bp and 500bp. This dataset doesn't follow a strict molecular clock.

The dataset already contained gene trees estimated from the sequence alignments with bootstrap values using RAxML version 7.3.5. The estimated gene trees already available in these datasets were used as the original gene trees in this study.

2.1.2 10-taxon

This dataset was previously studied in [4] and generated using SimPhy [29]. This dataset provides two different ILS conditions, the first having moderately high ILS (40% AD) and second having high ILS (84% AD). This dataset is very heterogeneous with a different species tree for each replicate and various genes having different rates of evolution. The average bootstrap support for this dataset ranged from 45% for the high ILS condition to 34% for the moderately high ILS condition. This dataset doesn't follow a strict molecular clock.

The dataset already contained gene trees estimated from the sequence alignments with bootstrap values using RAxML version 7.3.5. The estimated gene trees already available in these datasets were used as the original gene trees in this study.

2.1.3 15-taxon

This dataset was simulated and studied in [4]. This dataset has high ILS level (82% AD). This dataset is very homogeneous with each replicate having the same caterpillar species tree. Gene trees from this dataset

follow a strict molecular clock and differ from each other only due to the multi-species coalescent process. The ultrametric gene trees were simulated using McCoal [30] for this dataset. In the simulation process, sequences of length 1000bp were simulated down the gene trees and then 100bp were sampled to vary the number of sites. The average bootstrap support for the 100bp condition was 35%, whereas the average bootstrap support for the 1000bp condition was 70%.

The dataset already contained gene trees estimated from the sequence alignments with bootstrap values using RAxML version 7.3.5. The estimated gene trees already available in these datasets were used as the original gene trees in this study.

2.1.4 11-taxon

This dataset was initially developed for [31], and later studied in [2]. Between the two 11-taxon datasets available from [2], the dataset with more ILS was used in this thesis. The dataset used in this thesis had a low level of ILS (15% AD). This dataset doesn't follow a strict molecular clock and is very heterogeneous with substantial rate variation among the gene trees and species tree. 100 gene multiple sequence alignments were available and used in this thesis with a length of 500bp.

The dataset already contained gene trees estimated from the sequence alignments with bootstrap values using RAxML version 7.3.5. The estimated gene trees already available in these datasets were used as the original gene trees in this study.

2.1.5 SimPhy

This dataset was developed for [14] and simulated using SimPhy [29]. This dataset was simulated with 200 taxa, variable tree lengths (500K, 2M and 10M generation) and speciation rates ($1e-6$ and $1e-7$ per generation). The tree length and speciation rate affects the amount of ILS, with shorter lengths resulting in shorter branches and therefore higher levels of ILS. Speciation rate impacts whether speciation events tend to happen close to leaves ($1e-6$) or close to the root ($1e-7$).

In this study, the sequence alignments for the first 50 taxa were sampled from the original multiple sequence alignment. Then, sequences of length 50bp, 100bp, and 200bp were sampled from the new sequence alignments having 50 taxa. Then, gene trees were estimated on these alignments using FastTree [28] without bootstrap support values. FastTree is a maximum likelihood method for estimating gene trees from gene alignments. FastTree also produces local support values for each edge in the estimated gene tree. The different tree shapes used in the simulation (i.e., the tree length and the speciation rate) produced many model conditions with amount of ILS ranging from 9% to 69% AD. This dataset is very heterogeneous with

a different species tree for each replicate and various genes having different rates of evolution. The sequences in this dataset do not evolve under a strict molecular clock.

2.2 Design

In this experimental study, the WSB+WQMC pipeline was evaluated through a range of experiments on various simulated datasets that varied in many respects (sequence length of locus, number of taxa, number of genes, deviation from the molecular clock, and ILS level). The experiments were designed to answer the following questions.

- The effect of WSB+WQMC on gene tree estimation error.
- The effect of WSB+WQMC on species tree estimation error when species trees are computed using either ASTRAL2 or ASTRID.
- The effect of algorithmic parameters in the WSB+WQMC pipeline.
- The relative performance of WSB+CAML and WSB+WQMC with respect to gene tree accuracy and species tree accuracy.

The experimental study in this paper was broken into two phases, namely the training and the testing phases. The datasets were also separated into two disjoint sets as shown in Tables 2.1 and 2.2 for training and testing purposes.

2.2.1 Training

The training phase was required in this study to select the confidence value c to be used in the testing phase. In the training phase, the Mammalian-0.5X, Mammalian-1X and Mammalian-2X simulated datasets were analysed using different confidence values when running the WSB+WQMC pipeline. These datasets were chosen for the training phase because they provided a wide range of ILS levels (21%, 30%, and 50% AD), and were used in [4] to evaluate WSB+CAML. Based on the results on these training datasets, a single confidence value was chosen for the testing phase. Moreover, on these datasets, the effect of using two different binning thresholds (Binning 100% and Binning 75%) was also explored.

The average bin sizes computed by the binning step in WSB+WQMC on the SimPhy-10M1e-7 dataset with different binning threshold values (Binning 90% and Binning 95%) were also analyzed in the training phase. This was done to select a single binning threshold value t to be used for the SimPhy datasets in the testing phase.

For all the testing datasets except the SimPhy datasets, no training was done to select a binning threshold value and binning threshold 75% was used in the testing phase. This binning threshold value was used because the same binning threshold value was also used in [4] for running WSB+CAML on these datasets.

Mammalian

The WSB+WQMC pipeline on the Mammalian-0.5X, Mammalian-1X and Mammalian-2X datasets was run for different model conditions, as shown in Table 2.1. The input gene trees for the WSB+WQMC pipeline were BestML gene trees estimated using RAxML version 7.3.5 with 200 bootstrap replicates. For model conditions with 50 genes, the first 50 gene trees were sampled from 200 genes trees and used as input.

To explore the effect of confidence values, the WSB+WQMC pipeline was run multiple times once for each confidence value (0.0, 0.2, and 0.3) with binning threshold 75%. Note that the confidence value of 0.0 leads to no up-weighting of genes and outputs the same new gene tree for all input gene trees within a bin. Another experiment to analyze the effect of binning threshold values was conducted by running the WSB+WQMC pipeline on the training datasets with binning threshold values 75% and 100% and confidence value 0.2. Binning 100% puts all the gene trees into a single bin (also referred as one-bin in this study). The new gene trees produced by each run of the WSB+WQMC pipeline on each model condition were then used to compute a species tree using ASTRAL2 [14]. For analyzing the WSB+WQMC pipeline, the accuracy of the new gene trees computed by WSB+WQMC was compared to the accuracy of the original gene trees. Also, the accuracy of the ASTRAL2 species tree computed on new gene trees was compared to the accuracy of the ASTRAL2 species tree computed on original gene trees. More details on the evaluation of gene tree estimation and species tree estimation are provided in section 2.3.

SimPhy

The binning step in the WSB+WQMC pipeline was run on the SimPhy-10M1e-7 dataset having 50 taxa on model conditions having two different numbers of genes (100 and 200 genes) and three different alignment lengths (50bp, 100bp, and 200bp alignment length). For the model condition with 100 genes, the first 100 gene trees were sampled and used as input. For model conditions with sequence lengths 50bp and 100bp, the first 50 and 100 sites were sampled from the original alignment.

The input gene trees were estimated from the sequence alignments using FastTree 2.1 without bootstrap support values. In order to run the binning step in WSB+WQMC on this dataset, local support values computed by FastTree were used instead of bootstrap support values. It is important to note that the FastTree local support values are very different from bootstrap support values, and generally tend to be much higher than the bootstrap support values. To determine the binning threshold to be used in the

testing phase, average bin sizes after the binning step in WSB+WQMC on the SimPhy-10M1e-7 dataset were computed with binning thresholds 90% and 95% (as shown in Table 2.4). Only these very high binning threshold values were analyzed in the training phase as other lower binning threshold values would have resulted in singleton bins. It was found that binning threshold value 90% produced substantially smaller bins than 95%, making the latter a more preferable choice. Therefore, the binning threshold value of 95% was chosen for running WSB+WQMC and WSB+CAML for the SimPhy datasets in the testing phase.

2.2.2 Testing

Based on the results obtained for the training datasets (detailed in chapter 4), it was observed that the confidence values 0.2 and 0.3 were at par with each other and were a better choice than the confidence value 0.0. Specifically, the confidence values 0.2 and 0.3 were notably better than the confidence value 0.0 for gene tree estimation on all model conditions. Moreover, in terms of species tree estimation the confidence values 0.2 and 0.3 performed better in medium and high ILS conditions and were only worse in low ILS. Therefore, confidence value 0.2 was chosen as the only confidence value used in the rest of this study. For the Mammalian, 10-taxon, 15-taxon and 11-taxon datasets, the binning threshold value of 75% was chosen to be used in the testing phase. For the SimPhy datasets, the binning threshold value of 95% was chosen to be used in the testing phase.

In this phase, experiments were conducted to analyze the performance of the WSB+WQMC pipeline on gene tree and species tree estimation using ASTRAL2 and ASTRID on the remaining datasets. Moreover, the WSB+CAML pipeline was also run on a subset of data and compared with the WSB+WQMC pipeline. For species tree estimation, ASTRAL2 and ASTRID were again used to compute species trees from gene trees.

Mammalian

In this phase, the WSB+WQMC pipeline was run on the Mammalian-1X dataset having 250bp alignment length, and the results obtained were compared with results from running WSB+WQMC on the Mammalian-1X dataset having 500bp alignment length. Moreover, the WSB+CAML pipeline was compared to the WSB+WQMC pipeline on the Mammalian-1X dataset having 200 genes and 250bp alignment length. Both the pipelines were run using binning threshold 75%. The input gene trees for WSB+WQMC and WSB+CAML were BestML gene trees estimated from the gene alignments using RAxML version 7.3.5 with 200 bootstrap replicates. Supergene trees for each bin in the WSB-CAML pipeline were computed using BestML unpartitioned maximum likelihood analysis using RAxML version 8.2.7.

10-taxon

The WSB+WQMC pipeline was run on the 10-taxon H ILS and 10-taxon MH ILS datasets for various model conditions having different number of genes (shown in Table 2.2). For model conditions with 25, 50, and 100 genes, the first 25, 50, and 100 gene trees were sampled from 200 genes trees respectively and used as input.

For each model condition, the WSB+WQMC pipeline was run using binning threshold 75% and evaluated in terms of gene tree and species tree estimation. The WSB+CAML pipeline was also run on the 10-taxon MH ILS dataset having 100 genes and 100bp alignment length using binning threshold 75% and compared with WSB+WQMC. The input gene trees for both WSB+WQMC and WSB+CAML were BestML gene trees estimated from gene alignments using RAxML version 7.3.5 with 200 bootstrap replicates. Supergene trees for each bin in the WSB-CAML pipeline were computed using BestML fully partitioned maximum likelihood analysis using RAxML version 8.2.7.

15-taxon

The WSB+WQMC pipeline was run on the 15-taxon H ILS dataset for different model conditions varying in number of genes and number of sites, as shown in Table 2.2. For model conditions with 25, 50, 100, and 200 genes, the first 25, 50, 100, and 200 gene trees were sampled from 1000 genes trees and used as input. For the model condition with sequence length 100bp, the first 100 sites were sampled from the original alignment.

For each model condition, the WSB+WQMC pipeline was run using binning threshold 75%. The WSB+CAML pipeline was also run on the 15-taxon H ILS dataset having 100 genes and 100bp alignment length using binning threshold 75% and compared with WSB+WQMC. The input gene trees for both WSB+WQMC and WSB+CAML were BestML gene trees estimated from gene alignments using RAxML version 7.3.5 with 200 bootstrap replicates. Supergene trees for each bin in the WSB-CAML pipeline were computed using BestML fully partitioned maximum likelihood analysis using RAxML version 8.2.7.

11-taxon

The WSB+WQMC pipeline was run on the 11-taxon dataset under different model conditions, as shown in Table 2.2. The WSB+WQMC pipeline was run using binning threshold 75%. For model conditions with 25 and 50 genes, the first 25 and 50 gene trees were sampled from 100 genes trees respectively and used as input.

The WSB+CAML pipeline was also run on the 11-taxon dataset having 100 genes and 500bp alignment length using binning threshold 75% and compared with WSB+WQMC. The input gene trees for both WSB+WQMC and WSB+CAML were BestML gene trees estimated from gene alignments using RAxML

version 7.3.5 with 200 bootstrap replicates. Supergene trees for each bin in the WSB-CAML pipeline were computed using BestML fully partitioned maximum likelihood analysis using RAxML version 8.2.7.

SimPhy

The WSB+WQMC pipeline was run on the SimPhy datasets under various model conditions, as shown in Table 2.2. For model conditions with 25, 50, and 100 genes, the first 25, 50, and 100 gene trees were sampled respectively and used as input. For model conditions with sequence lengths 50bp and 100bp, the first 50 and 100 sites were sampled from the original alignment.

The input gene trees were estimated from the gene alignments using FastTree 2.1 without bootstrap values. In order to do the binning step in the WSB+WQMC and WSB+CAML pipelines, the FastTree local support values were used instead of bootstrap support values. Apart from using the FastTree local support values instead of bootstrap support values, everything else was kept unchanged for running the WSB+WQMC and WSB+CAML pipelines. Based on the results obtained in the training phase, the binning threshold value of 95% was chosen to be use in the testing phase.

The WSB+CAML pipeline was also run on the SimPhy datasets with model conditions having 100 genes and 100bp alignment length and compared with WSB+WQMC. Supergene trees for each bin in the WSB-CAML pipeline were computed using BestML unpartitioned maximum likelihood analysis using FastTree version 2.1.

2.3 Methods

This section describes in detail the software and methods used for new gene tree and species tree estimation in this experimental study.

2.3.1 New gene tree estimation

The WSB+WQMC pipeline computes a set of weighted quartets for each gene within a bin. WQMC is then run with the unique weighted quartet set as input to compute a new gene tree. Since the binning step computes disjoint bins, a new gene tree is obtained for each gene. The WSB+CAML pipeline on the other hand computes a supergene alignment for each bin. Maximum likelihood gene tree estimation method (RAxML or FastTree) is then run on each bin to compute a supergene tree for that bin. Finally, for each gene within a bin, the supergene tree computed for that bin is considered as the new gene tree.

2.3.2 Species tree estimation

To compute species trees from gene trees, summary methods ASTRID 1.1 [15] and ASTRAL2 4.7.2 [14] were used for the basic unbinned, WSB+WQMC, and WSB+CAML pipelines.

ASTRAL2 in the default mode uses the bipartitions from the input gene trees to constraint the search space for the output species tree. To expand this search space, ASTRAL2 allows a set of extra bipartitions to be given as input. In this thesis, ASTRAL2 was run on three different sets of gene trees: the original set of gene trees, the newly estimated WSB+WQMC gene trees, and the newly estimated WSB+CAML gene trees. To enable a fair comparison between these three ways of computing species trees, the same set of extra bipartitions for ASTRAL2 was used. This set of extra bipartitions contained bipartitions from all the trees listed below.

- Original gene trees.
- New gene trees computed by running WSB+WQMC.
- Species tree computed by running ASTRID on the original gene trees.
- Species tree computed by running ASTRID on the new gene trees computed by WSB+WQMC.
- Species tree computed by running ASTRAL2 in default mode on the original gene trees.

2.3.3 Measurement

The new gene trees computed by both WSB+WQMC and WSB+CAML were evaluated by calculating gene tree estimation error. Gene tree estimation error was computed by measuring the average missing branch rate (FN rate) of the new gene trees with respect to the true gene trees. Gene tree estimation error for the basic unbinned pipeline was calculated by measuring the average missing branch rate of the original gene trees with respect to the true gene trees.

Similarly, species tree estimation error was computed for both WSB+WQMC and WSB+CAML by measuring the average missing branch rate of the estimated species tree on new gene trees with respect to the true species tree. For the basic unbinned pipeline, species tree estimation error was calculated by measuring the average missing branch rate of the estimated species tree on original gene trees with respect to the true species tree.

For both gene tree estimation error and species tree estimation error, the average was computed over 10 replicates in each dataset.

2.4 Figures and Tables

Dataset	Taxa	AD	Genes	Sites	Reps	Bootstraps	Ref
Mammalian-0.5X	37	50	50,200	500	10	Available	[3]
Mammalian-1X	37	30	50,200	500	10	Available	[3]
Mammalian-2X	37	21	50,200	500	10	Available	[3]

Table 2.1: **Empirical statistics of simulated datasets used in training phase.** The ILS level is measured by AD%, which is the percentage of missing branches between true gene trees and the species tree. Bootstraps column indicates whether bootstrap support values were available or not on estimated gene trees.

Dataset	Taxa	AD	Genes	Sites	Reps	Bootstraps	Ref
Mammalian-1X	37	30	50,200	250	10	Available	[3]
10-taxon H ILS	10	84	25,50,100,200	100	10	Available	[4]
10-taxon MH ILS	10	40	25,50,100,200	100	10	Available	[4]
15-taxon H ILS	15	82	25,50,100,200,1000	100,1000	10	Available	[4]
11-taxon L ILS	11	15	25,50,100	500	10	Available	[2]
SimPhy-10M1e-7	50	9	25,50,100,200	50,100,200	10	NA	[14]
SimPhy-10M1e-6	50	21	25,50,100,200	50,100,200	10	NA	[14]
SimPhy-2M1e-6	50	30	25,50,100,200	50,100,200	10	NA	[14]
SimPhy-2M1e-7	50	34	25,50,100,200	50,100,200	10	NA	[14]
SimPhy-500K1e-7	50	68	100	100	10	NA	[14]
SimPhy-500K1e-6	50	69	100	100	10	NA	[14]

Table 2.2: **Empirical statistics of simulated datasets used in testing phase.** The ILS level is measured by AD%, which is the percentage of missing branches between true gene trees and the species tree. Bootstraps column indicates whether bootstrap support values were available or not on estimated gene trees.

ILS Level	Range of AD%
Low (L)	< 22%
Medium (M)	22% – 30%
Moderately high (MH)	31% – 40%
High (H)	> 40%

Table 2.3: **Different categories of ILS level used in the study.** The ILS level is measured by AD%, which is the percentage of missing branches between true gene trees and the species tree.

Genes	Sites	Avg bin-size ($t = 90\%$)	Avg bin-size ($t = 95\%$)
200	50	2.23	6.12
100	50	2.09	5.42
200	100	2.85	7.64
100	100	2.55	6.93
200	200	2.65	7.30
100	200	2.47	6.29

Table 2.4: **Average bin-sizes for the SimPhy-10M1e-7 (9% AD) dataset for different model conditions and binning thresholds t .** Statistical Binning was used to compute bins with different threshold values t . FastTree local support values were used instead of bootstrap support values for running Statistical Binning. The average was taken over 10 replicates.

Chapter 3

Methods and Commands

3.1 Data availability

All datasets and supporting online materials are available at <http://goo.gl/zR48CT>. A github repository containing all the source code used in the experiments can be found at https://github.com/agupta0905/improving_genes.

3.2 Commands

Gene tree estimation

RAxML 7.3.5 was used in [2–4] for the Mammalian, 10-taxon, 15-taxon and 11-taxon datasets to estimate each original gene tree from its sequence alignment with bootstrap support with the following commands.

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

The following command was used for bootstrapping:

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

FastTree 2.1 was used to estimate each original gene tree from its sequence alignment for the SimPhy dataset using the following command.

```
FastTree -nt -gtr -gamma < [input_alignment] > [output_genetree]
```

WQMC was used to estimate each new gene tree from its weighted quartet file in the WSB+WQMC pipeline using the following command.

```
max-cut-tree qrtt=[weight_quartet_file] weights=on otre=[output_new_genetree]
```

Supergene tree estimation

RAxML 8.2.7 was used to estimate the supergene tree from the supergene alignment for the 10-taxon, 15-taxon and 11-taxon datasets in fully partitioned mode with the following command.

```
raxmlHPC-SSE3 -m GTRGAMMA -s [input_alignment] -n [output_name] -N 20  
-M -q [partition_file] -p [random_number]
```

RAxML 8.2.7 was used to estimate the supergene tree from the supergene alignment for the mammalian dataset in unpartitioned mode with the following command.

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

FastTree 2.1 was used to estimate the supergene tree from the supergene alignment for the SimPhy dataset in unpartitioned mode with the following command.

```
FastTree -nt -gtr -gamma < [input_alignment] > [output_genetree]
```

Species tree estimation

ASTRID 1.1 was used to estimate the species tree from original gene trees, WSB+WQMC gene trees, and WSB+CAML gene trees with the following command.

```
ASTRID -i [input_gene_trees] -o [output_species_tree]
```

ASTRAL2 4.7.2 was used in default mode to estimate the species tree from original gene trees with the following command.

```
java -jar astral.4.7.12.jar -i [input_gene_trees] -o [output_species_tree]
```

ASTRAL2 4.7.2 was used with a set of extra bipartitions to estimate the species tree from original gene trees, WSB+WQMC gene trees, and WSB+CAML gene trees with the following command.

```
java -jar astral.4.7.12.jar -i [input_gene_trees] -o [output_species_tree]  
-e [extra_gene_trees]
```

Chapter 4

Results

4.1 Training

Impact of confidence value c and different binning thresholds on gene tree estimation error

Gene tree estimation results on the Mammalian-0.5X, Mammalian-1X, and Mammalian-2X datasets with 200 genes and 500bp alignment length by running the WSB+WQMC pipeline with varying confidence values and binning threshold 75% are shown in Figure 4.1. It was observed that the new gene trees computed by WSB+WQMC with confidence value 0.0 were consistently worse than WSB+WQMC with confidence values 0.2 and 0.3 among all ILS levels. WSB+WQMC with confidence value 0.2 and 0.3 produced similar results and were very close to each other. Compared to the original gene trees, the WSB+WQMC pipeline with confidence value 0.2 and 0.3 produced worse gene trees on high ILS Mammalian-0.5X (50% AD), similar gene trees on medium ILS Mammalian-1X (30% AD), and better gene trees on low ILS Mammalian-0.5X (21% AD).

Figure 4.2 shows gene tree estimation results on the Mammalian-0.5X, Mammalian-1X, and Mammalian-2X datasets with 200 genes and 500bp alignment length by running the WSB+WQMC pipeline using binning threshold 75% and binning threshold 100% (one-bin) with confidence value 0.2. It was observed that binning 75% outperformed binning 100% in medium and high ILS conditions, with binning 100% being much worse and computing highly inaccurate gene trees for high ILS. In low ILS conditions, binning 100% performed better than binning 75%. Compared to the original gene trees, WSB+WQMC computed better gene trees in low ILS conditions, slightly worse gene trees in medium ILS, and worse gene trees in high ILS conditions.

We observed that the new gene trees got better on the Mammalian-2X dataset (21% AD), were similar to the original gene trees on the Mammalian-1X dataset (30% AD), and got worse in the Mammalian-0.5X

dataset (50% AD) for both binning thresholds. Also, it was observed that between binning 100% and binning 75%, binning 100% got higher magnitudes of change with respect to the original gene trees (substantially worse in high ILS and more accurate in low ILS).

Impact of confidence value c and different binning thresholds on species tree estimation error

ASTRAL2 species tree estimation results on the Mammalian-0.5X, Mammalian-1X, and Mammalian-2X datasets with 200 genes and 500bp alignment by running the WSB+WQMC pipeline with varying confidence values and binning threshold 75% are shown in Figure 4.3. It was observed that WSB+WQMC with confidence value 0.0 was worse than confidence values 0.2 and 0.3 for the Mammalian-0.5X (50% AD) and Mammalian-1X (30% AD) datasets but better for the Mammalian-2X (21% AD) dataset. Compared to each other, confidence values 0.2 and 0.3 were very close to each other. For all ILS levels and confidence values, it was observed that WSB+WQMC failed to compute a better species tree than the species tree on original gene trees.

Figure 4.4 shows ASTRAL2 species tree estimation results on the Mammalian-0.5X, Mammalian-1X, and Mammalian-2X datasets with 200 genes and 500bp alignment length by running the WSB+WQMC pipeline using binning threshold 75% and binning threshold 100% (one-bin) with confidence value 0.2. It was observed that binning 75% outperformed binning 100% in medium and high ILS conditions, with binning 100% being much worse and computing highly inaccurate species trees in high ILS conditions. In low ILS conditions, binning 100% performed better than binning 75%. Overall, both binning thresholds used for running WSB+WQMC failed to compute a more accurate species tree compared to the species tree computed on the original gene trees.

4.2 Testing

Impact of WSB+WQMC on gene tree estimation error

In the following section, the accuracy of new gene trees computed by running the WSB+WQMC pipeline is compared to the accuracy of original gene trees.

Gene tree estimation results after running WSB+WQMC on all datasets and all model conditions listed in Tables 2.2 and 2.1, with an exception of the SimPhy-500K1e-7 and SimPhy-500K1e-6 datasets, are shown in Figure 4.5. In general, in low ILS levels the WSB+WQMC gene trees got more accurate than the original gene trees. With increasing ILS, the improvement in the accuracy of WSB+WQMC gene trees over the

original gene trees decreased and finally WSB+WQMC gene trees got less accurate in high ILS conditions. Results show that WSB+WQMC improves gene tree accuracy for datasets having low level of ILS and decreases accuracy of gene trees for datasets having high levels of ILS. For medium and moderately high ILS levels, the relative performance of WSB+WQMC compared to the basic unbinned pipeline was dependent on the dataset and model condition itself. The Mammalian-1X (30% AD) and SimPhy 2M1e-7 (34% AD) datasets were better off using the WSB+WQMC pipeline on some model conditions but worse for others. On the other hand, on the 10-taxon MH ILS dataset (40% AD) it was better to use WSB+WQMC than the basic unbinned pipeline on all model conditions.

This strong relationship between the ILS level and the improvement in gene tree estimation error is also seen in Figure 4.6, where WSB+WQMC substantially improves the accuracy of gene trees for the 10-taxon MH ILS dataset (40% AD, 200 genes and 100bp) but makes the gene trees worse for the 10-taxon H ILS dataset (84% AD, 200 genes and 100bp). The degree of improvement in the accuracy of gene trees was also seen to be directly related to the ILS level. Results on the 11-taxon L ILS dataset shown in Figure 4.7 and SimPhy datasets (except 500K1e-6 and 500K1e-7) shown in Figure 4.8 clearly show huge improvements in gene tree accuracy for the 11-taxon L ILS (15% AD), 10M1e-7 (9% AD), and 10M1e-6 datasets (21% AD), whereas the improvements in gene tree accuracy for 2M1e-6 (30% AD) and 2M1e-7 (34 % AD) datasets having medium and moderately high ILS levels were lower in magnitude.

The effect of varying alignment length on the accuracy of WSB+WQMC gene trees was also studied for the Mammalian-1X and 15-taxon H ILS datasets, as shown in Figures 4.9 and 4.10. WSB+WQMC was run on the Mammalian-1X dataset (30% AD) having 200 genes with two model conditions having 250bp and 500bp alignment lengths. For the 15-taxon H ILS dataset (82% AD), WSB+WQMC was run on two different model conditions having 100bp and 1000bp alignment lengths. On the Mammalian-1X dataset, shorter alignment length improved the accuracy of gene trees substantially while longer alignment length computed worse gene trees compared to the original gene trees. For the 15-taxon H ILS dataset both alignment lengths computed better WSB+WQMC gene trees than the original gene trees. However, in the model condition having 1000bp alignment length, the accuracy decreased by a greater factor compared to the 100bp alignment length model condition. These results suggest that the benefit of using WSB+WQMC to compute gene trees compared to using the basic unbinned pipeline was greater for model conditions having smaller alignment lengths.

Comparison of WSB+WQMC and WSB+CAML in terms of gene tree estimation error

When comparing WSB+WQMC with WSB+CAML on the 11-taxon L ILS (15% AD), Mammalian-1X (30% AD), 10-taxon MH ILS (40% AD), and 15-taxon H ILS (82% AD) datasets (as shown in Figure 4.11), WSB+WQMC computed better gene trees than WSB+CAML for all datasets and all ILS levels in the experiment. Both pipelines computed better gene trees than the original gene trees for the 11-taxon L ILS, Mammalian-1X, and 10-taxon MH ILS datasets but worse gene trees for the 15-taxon H ILS dataset.

In contrast, when comparing WSB+WQMC with WSB+CAML on the SimPhy datasets having speciation rates $1e-7$ and $1e-6$ (shown in Figure 4.12), WSB+CAML computed better gene trees for low ILS and medium ILS datasets (i.e., 10M1e-6, 10M1e-7, and 2M1e-6) and worse gene trees for high ILS datasets (i.e., 500K1e-6 and 500K1e-7). For moderately high ILS datasets (i.e., 2M1e-7), both WSB+WQMC and WSB+CAML performed equally well in estimating gene trees. Compared to the original gene trees, both pipelines computed better gene trees in low, medium, and moderately high ILS levels. In high ILS levels, both pipelines computed worse gene trees compared to the original gene trees.

In terms of gene tree estimation, we didn't find a clear winner between WSB+CAML and WSB+WQMC, as their relative performance was dependent on the ILS level and the dataset itself. It was found that WSB+WQMC computed better gene trees than WSB+CAML in high ILS conditions. In moderately high ILS levels, WSB+WQMC performed better than WSB+CAML on the 10-taxon MH ILS (40% AD) dataset but similar to WSB+CAML on the SimPhy 2M1e-7 (34% AD) dataset. In medium and low ILS conditions, WSB+WQMC computed better gene trees than WSB+CAML on the 11-taxon L ILS (15% AD) and Mammalian-1X (30% AD) but worse gene trees on the SimPhy 10M1e-7 (9% AD), 10M1e-7(21% AD), and 2M1e-6 (30% AD) datasets.

Impact of WSB+WQMC on species tree estimation error

In the following section, the accuracy of species trees computed by running the WSB+WQMC pipeline is compared to the accuracy of species trees computed by running the basic unbinned pipeline.

ASTRAL2 species tree estimation results on all datasets and all model conditions listed in Tables 2.2 and 2.1, with an exception of the SimPhy-500K1e-7 and SimPhy-500K1e-6 datasets, are shown in Figure 4.13. In general, in low ILS levels the ASTRAL2 species tree on WSB+WQMC gene trees got more accurate than the ASTRAL2 species tree on original gene trees. With increasing ILS, the improvement in accuracy of ASTRAL2 species tree using the WSB+WQMC pipeline decreased and finally the ASTRAL2 species tree computed on WSB+WQMC gene trees got less accurate in high ILS conditions. Although WSB+WQMC

computed more accurate ASTRAL2 species trees on most model conditions having low ILS, there were a few exceptions. There was no clear pattern among these low ILS model conditions where WSB+WQMC was computing less accurate ASTRAL2 species trees. In medium and moderately high ILS levels, most model conditions benefited by using the WSB+WQMC pipeline, but there were some model conditions among these datasets where WSB+WQMC performed poorly. The results showed another interesting trend with the improvement in accuracy of ASTRAL2 species tree using the WSB+WQMC pipeline for some model conditions on the 15-taxon H ILS (82% AD) and 10-taxon (84% AD) datasets having high levels of ILS. Again, there was no clear pattern among the model conditions having high ILS and more accurate ASTRAL2 species trees computed using the WSB+WQMC pipeline.

This weak correlation between the ILS level and the improvement in species tree estimation error using the WSB+WQMC pipeline was found throughout the experiments. Figure 4.14 shows results on the 11-taxon L ILS (15% AD) dataset, where WSB+WQMC improved the accuracy of the ASTRID species tree but had no change for the ASTRAL2 species tree. ASTRID and ASTRAL2 species trees computed using WSB+WQMC gene trees on low, medium, and moderately high ILS SimPhy datasets (i.e., $10M1e-7$, $10M1e-6$, $2M1e-6$, and $2M1e-7$) were seen to be substantially better than the species trees computed on original gene trees (see Figure 4.15). Results on the 10-taxon MH ILS (40% AD) and 10-taxon H ILS (82%) datasets (shown in Figure 4.16) show that ASTRID and ASTRAL2 species trees computed on WSB+WQMC gene trees were less accurate than the species trees computed on original gene trees. In contrast, WSB+WQMC was able to reduce species tree estimation error for the 15-taxon H ILS (82% AD) dataset (shown in Figure 4.17), despite having high ILS. Both ASTRID and ASTRAL2 species trees got more accurate for the 15-taxon dataset with 100bp alignment length model condition but only the ASTRID species tree was improved for the 15-taxon 1000bp model condition. When compared to the unbinned pipeline, it was also observed that on the 15-taxon dataset WSB+WQMC performed better in the shorter alignment model condition (100bp) than the longer alignment model condition (1000bp), with greater magnitude of improvements in the former case. Figure 4.9 shows the impact of alignment length on the accuracy of species trees computed by WSB+WQMC on the Mammalian-1X (30% AD) dataset. In this experiment, alignment lengths of 250bp and 500bp were used. It was found that for both model conditions, ASTRAL2 and ASTRID species trees on WSB+WQMC gene trees were worse than the species trees on original gene trees. Varying alignment length didn't seem to affect the performance of WSB+WQMC on this dataset.

Comparison of WSB+WQMC and WSB+CAML in terms of species tree estimation error

Figures 4.19 and 4.20 show the species tree estimation results by running the WSB+WQMC and WSB+CAML pipelines on the 11-taxon (15% AD), Mammalian-1X (30% AD), 10-taxon MH ILS (40% AD), and 15-taxon H ILS (82% AD) datasets. It was found that on moderately high and high ILS datasets (i.e., 10-taxon MH ILS and 15-taxon H ILS), WSB+WQMC computed more accurate ASTRAL2 species trees than WSB+CAML. Also, it was found that WSB+WQMC computed better ASTRID species trees than WSB+CAML on the 10-taxon dataset but similar species trees on the 15-taxon dataset. However, on low and medium ILS level datasets (i.e., 11-taxon and Mammalian-1X datasets), WSB-CAML outperformed WSB-WQMC in computing more accurate species trees using both ASTRID and ASTRAL2. It was found that the WSB+WQMC pipeline computed similar or better ASTRAL2 species tree than the basic unbinned pipeline for the 11-taxon L ILS and 15-taxon H ILS datasets, whereas it computed better ASTRID species tree for all datasets used in this experiment except the Mammalian-1X dataset.

Figures 4.21 and 4.22 show the species tree estimation results by running the WSB+WQMC and WSB+CAML pipelines on the SimPhy datasets having speciation rates $1e-7$ and $1e-6$. WSB+CAML computed better species trees for low and medium ILS datasets (i.e., $10M1e-7$, $10M1e-6$, and $2M1e-6$) using both ASTRID and ASTRAL2. For moderately high and high ILS datasets (i.e., $2M1e-7$, $500K1e-7$, and $500K1e-6$), WSB+WQMC was found to be at least as accurate as WSB+CAML in computing species trees using both ASTRID and ASTRAL2. Moreover, the ASTRID trees computed by WSB+WQMC were found to be substantially better than the ASTRID trees computed by WSB+CAML on the SimPhy- $2M1e-7$ and SimPhy- $500K1e-6$ datasets. Both pipelines performed better than the basic unbinned pipeline in computing species trees for low and medium ILS levels irrespective of the summary method. In moderately high ILS conditions, WSB+WQMC computed better species trees than the basic unbinned pipeline for both ASTRID and ASTRAL2. However, WSB+CAML computed better ASTRAL2 species trees but worse ASTRID species trees than the basic unbinned pipeline on moderately high ILS datasets. For high ILS datasets, both pipelines computed worse species trees than the species trees on original gene trees for both ASTRID and ASTRAL2.

In summary, it was found that WSB+WQMC computed similar or better species trees than WSB+CAML in moderately high and high ILS conditions irrespective of the summary method. However, in low and medium ILS conditions, WSB+CAML outperformed WSB+WQMC in computing the species trees using both ASTRID and ASTRAL2.

4.3 Figures and Tables

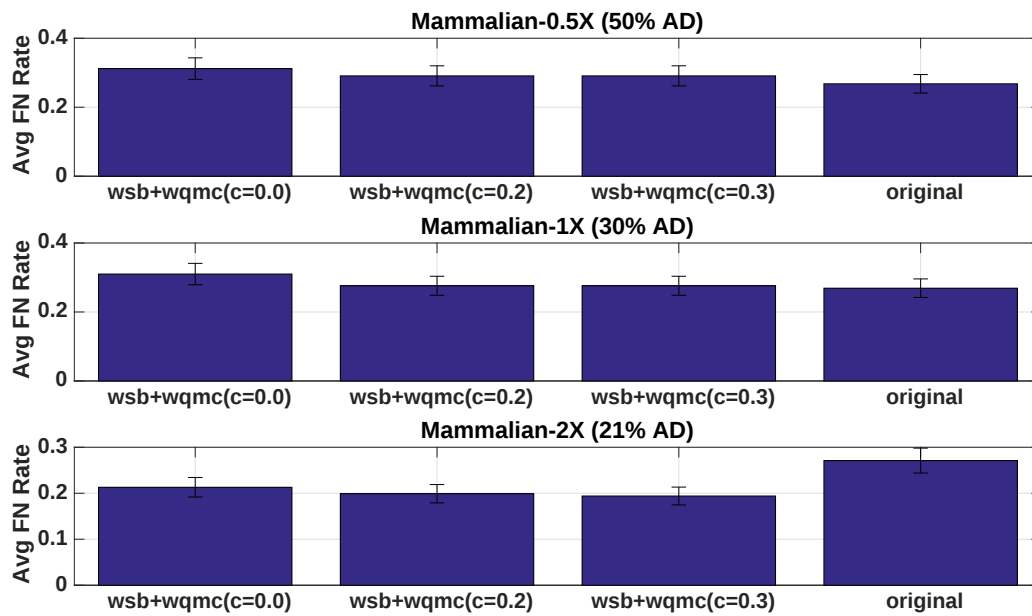


Figure 4.1: **Gene tree estimation error on the Mammalian datasets for running WSB+WQMC with different confidence values and binning threshold 75%.** We show results on the Mammalian-0.5X (50% AD), Mammalian-1X (30% AD), and Mammalian-2X (21% AD) datasets with 200 genes and 500bp alignment length. The x-axis shows three different ways of running WSB+WQMC by varying confidence value c (0.0, 0.2, and 0.3). The accuracy of WSB+WQMC gene trees is also compared to the accuracy of original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

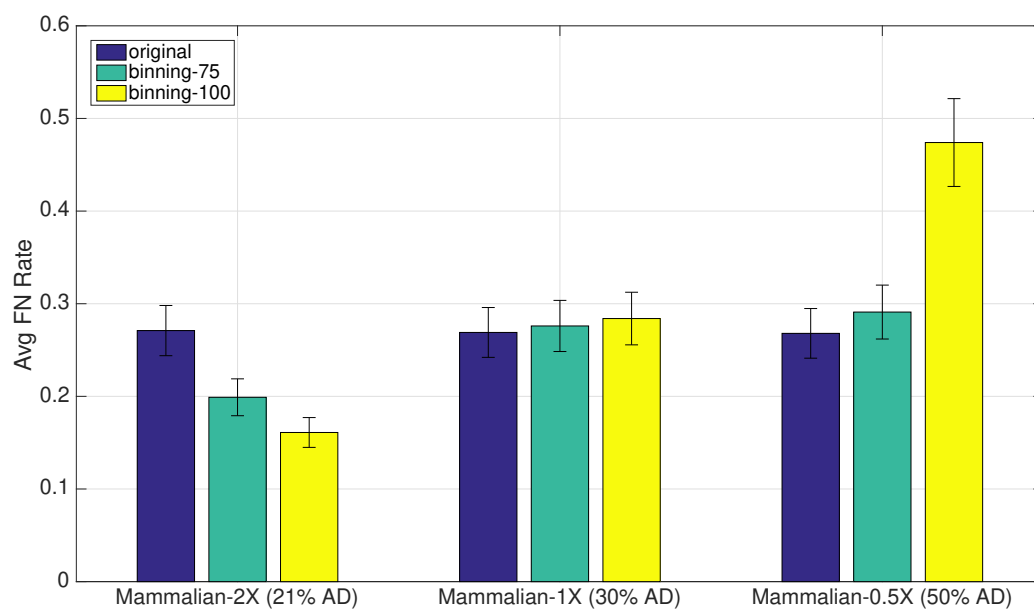


Figure 4.2: **Gene tree estimation error on the Mammalian datasets for running WSB+WQMC with binning threshold 75% and binning threshold 100% using confidence value 0.2.** We show results on the Mammalian-0.5X (50% AD), Mammalian-1X (30% AD), and Mammalian-2X (21% AD) datasets with 200 genes and 500bp alignment length. The accuracy of WSB+WQMC gene trees is also compared to the accuracy of original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

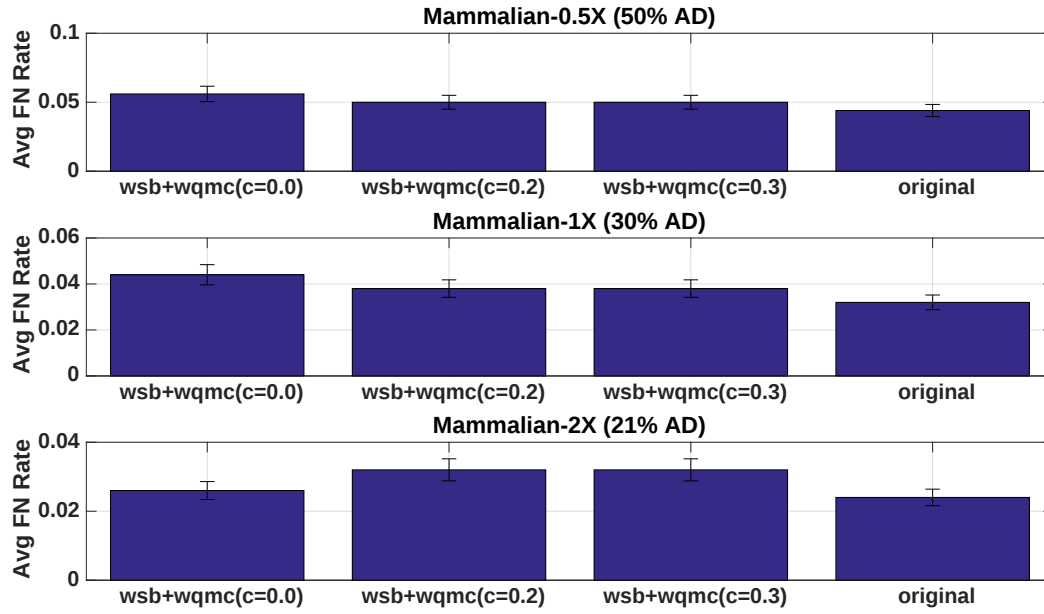


Figure 4.3: **ASTRAL2** species tree estimation error on the Mammalian datasets for running **WSB+WQMC** with different confidence values and binning threshold 75%. We show results on the Mammalian-0.5X (50% AD), Mammalian-1X (30% AD), and Mammalian-2X (21% AD) datasets with 200 genes and 500bp alignment length. The x-axis shows three different ways of running WSB+WQMC by varying confidence value c (0.0, 0.2, and 0.3). The accuracy of species tree on WSB+WQMC gene trees is also compared to the accuracy of species tree on original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

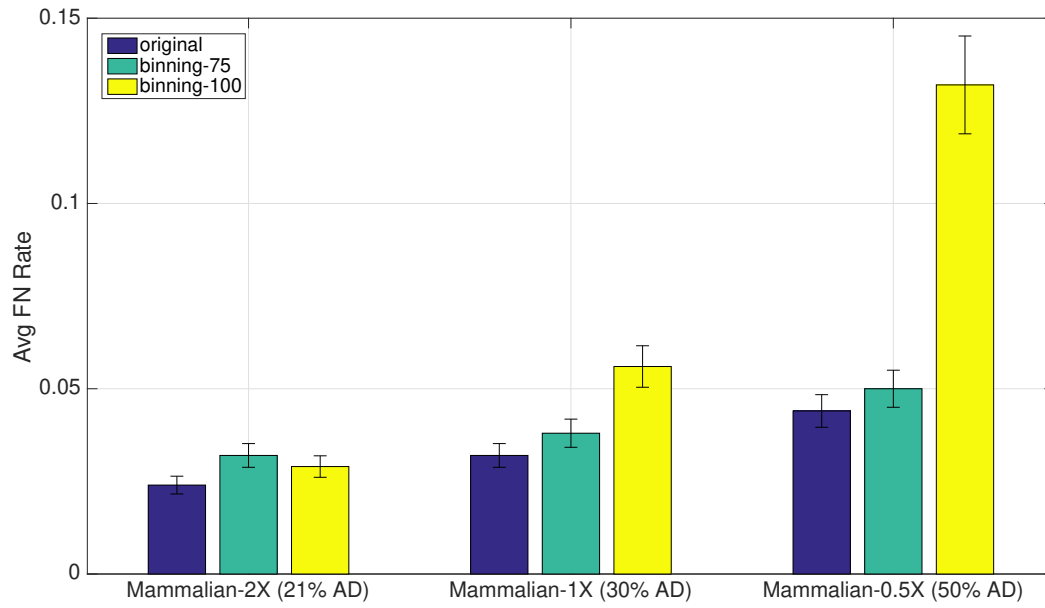


Figure 4.4: **ASTRAL2 species tree estimation error on the Mammalian datasets for running WSB+WQMC with binning threshold 75% and binning threshold 100% using confidence value 0.2.** We show results on the Mammalian-0.5X (50% AD), Mammalian-1X (30% AD), and Mammalian-2X (21% AD) datasets with 200 genes and 500bp alignment length. The accuracy of species tree on WSB+WQMC gene trees is also compared to the accuracy of species tree on original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

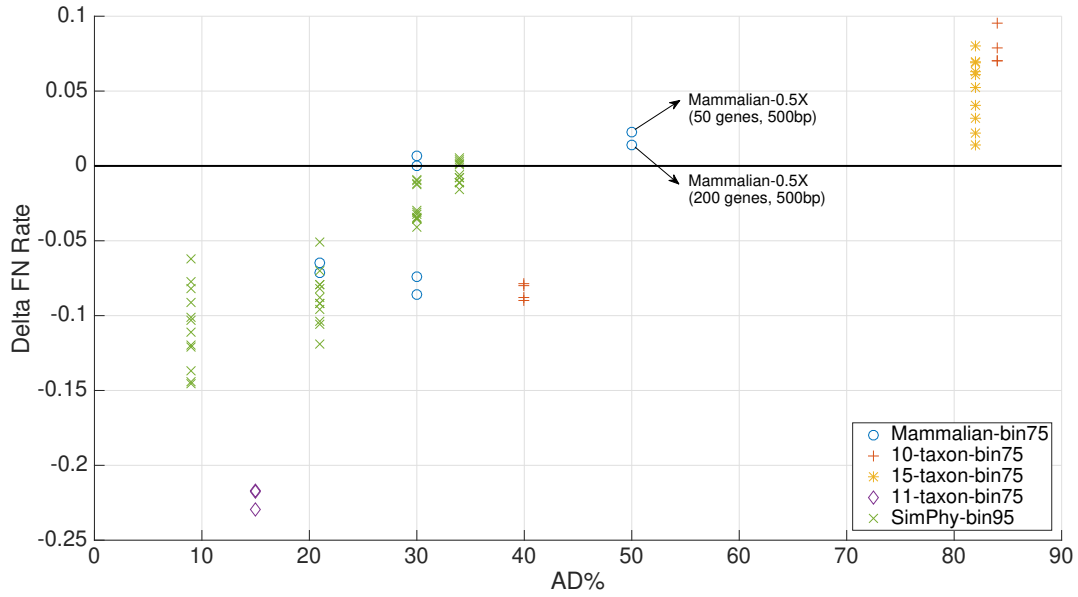


Figure 4.5: **Gene tree estimation error results on running WSB+WQMC on different model conditions of datasets shown in Tables 2.1 and 2.2.** We show results on the Mammalian-0.5X (50% AD), Mammalian-1X (30% AD), Mammalian-2X (21% AD), SimPhy-10M1e-7 (9% AD), SimPhy-10M1e-6 (21% AD), SimPhy-2M1e-6 (30% AD), SimPhy-2M1e-7 (34% AD), 10-taxon MH ILS (40% AD), 10-taxon H ILS (84% AD), 15-taxon H ILS (82% AD), and 11-taxon L ILS (15% AD) datasets with different number of genes and alignment lengths, as shown in Tables 2.1 and 2.2. The ILS level is varied on the x-axis with increasing ILS from left to right. Delta FN rate was computed by subtracting average FN rate of original gene trees from average FN rate of WSB+WQMC gene trees. Average FN rates were computed over 10 replicates for each dataset and model condition. Each point in the plot corresponds to delta FN rate on some model condition of the corresponding dataset. WSB+WQMC was run on these datasets with confidence value 0.2 and binning thresholds indicated in the legend.

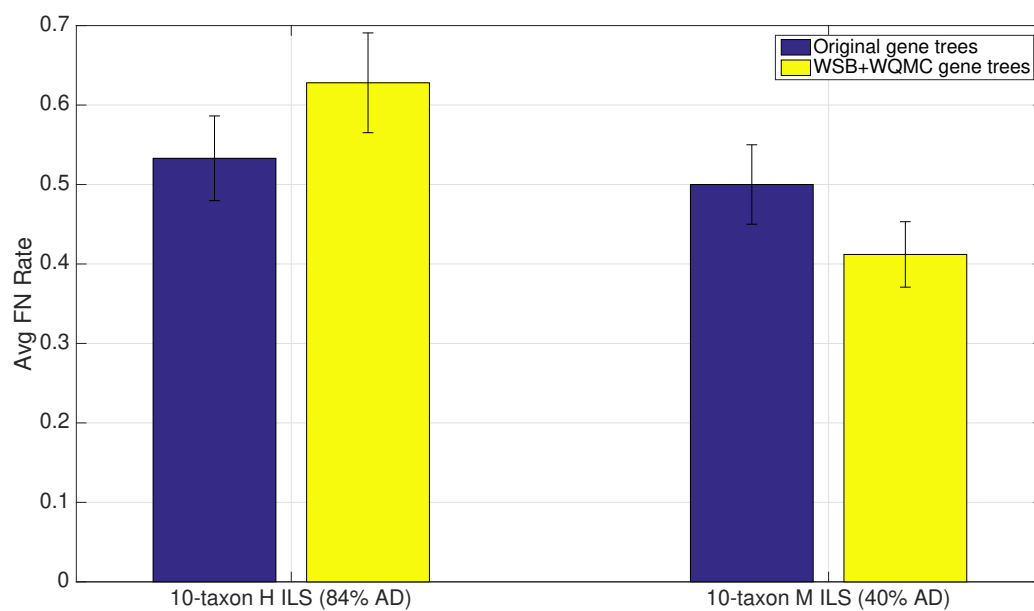


Figure 4.6: **Gene tree estimation error on 10-taxon datasets for running WSB+WQMC.** We show results on the 10-taxon MH ILS (40% AD) and 10-taxon H ILS (84% AD) with 200 genes and 100bp alignment length. WSB+WQMC was run using binning threshold 75% and confidence value 0.2. The accuracy of WSB+WQMC gene trees is also compared to the accuracy of original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

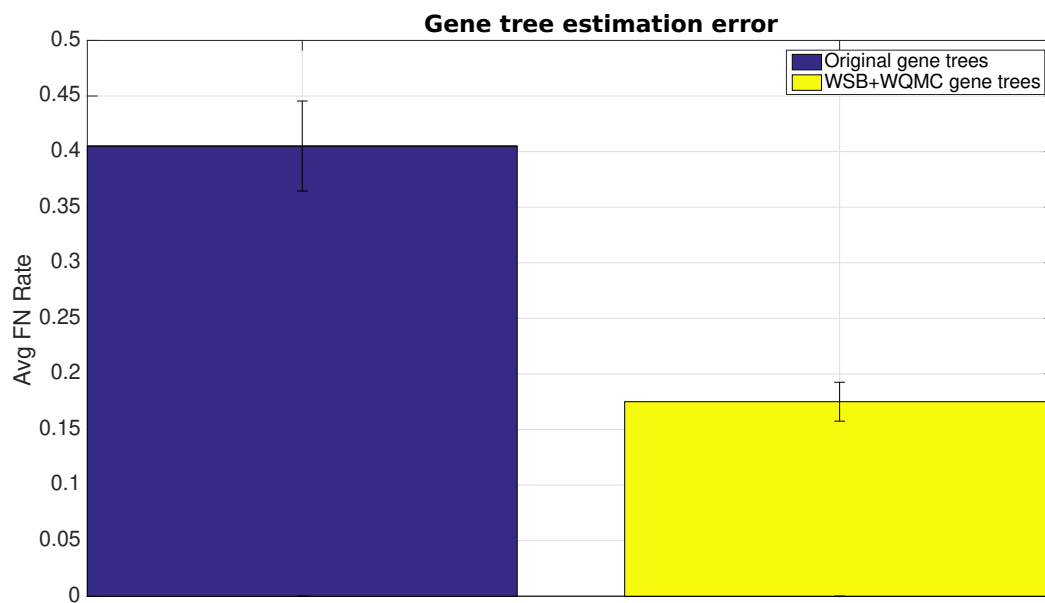


Figure 4.7: **Gene tree estimation error on 11-taxon dataset for running WSB+WQMC.** We show results on the 11-taxon L ILS (15% AD) dataset was used having 100 genes and 500bp alignment length. WSB+WQMC was run using binning threshold 75% and confidence value 0.2. The accuracy of WSB+WQMC gene trees is also compared to the accuracy of original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

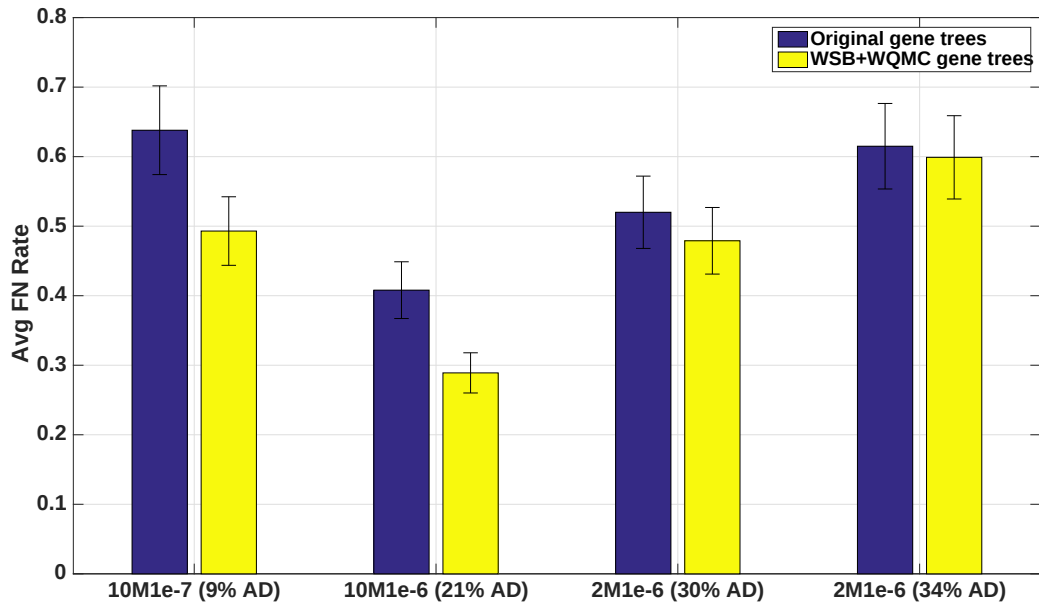


Figure 4.8: **Gene tree estimation error on SimPhy datasets for running WSB+WQMC.** We show results on the 10M1e-7 (9% AD), 10M1e-6 (21% AD), 2M1e-6 (30% AD), and 2M1e-7 (34% AD) datasets with 200 genes and 200bp alignment length. WSB+WQMC was run using binning threshold 95% and confidence value 0.2. The accuracy of WSB+WQMC gene trees is also compared to the accuracy of original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

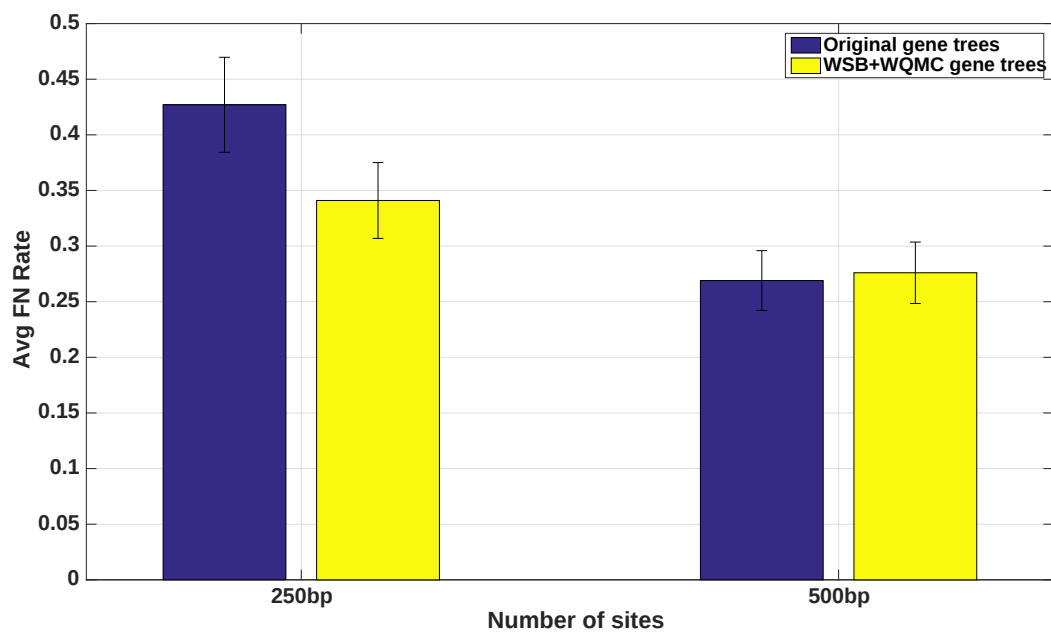


Figure 4.9: **Gene tree estimation error on the Mammalian-1X dataset with two different alignment lengths for running WSB+WQMC.** We show results on the Mammalian-1X (30% AD) dataset was used having 200 genes and two different alignment lengths (250bp and 500bp). WSB+WQMC was run using binning threshold 75% and confidence value 0.2. The accuracy of WSB+WQMC gene trees is also compared to the accuracy of original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

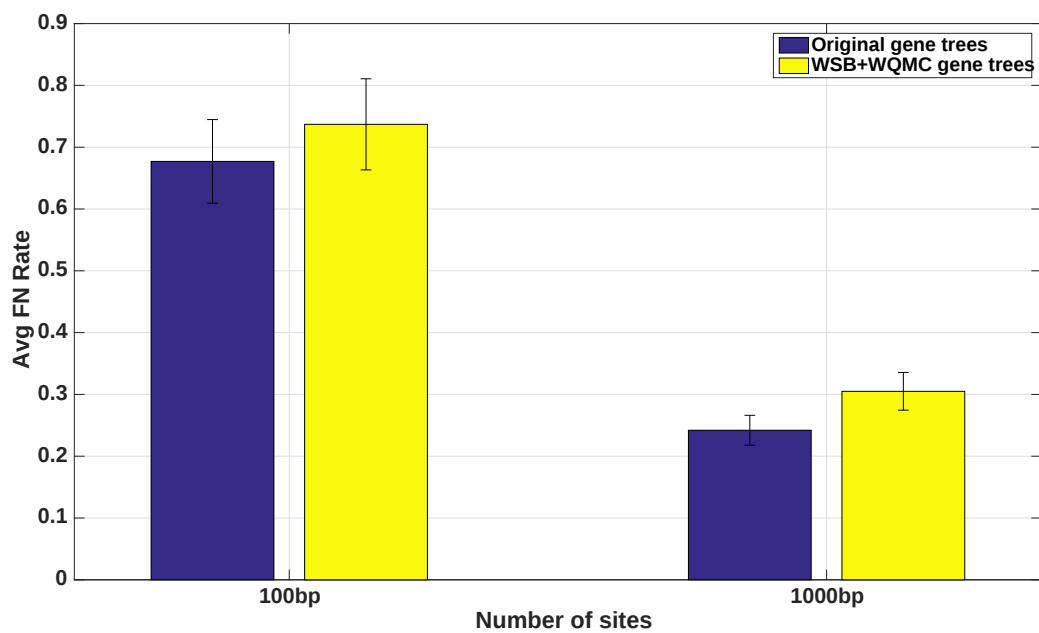


Figure 4.10: **Gene tree estimation error on 15-taxon H ILS dataset with two different alignment lengths for running WSB+WQMC.** We show results on the 15-taxon H ILS (82% AD) dataset was used having 1000 genes and two different alignment lengths (100bp and 1000bp). WSB+WQMC was run using binning threshold 75% and confidence value 0.2. The accuracy of WSB+WQMC gene trees is also compared to the accuracy of original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

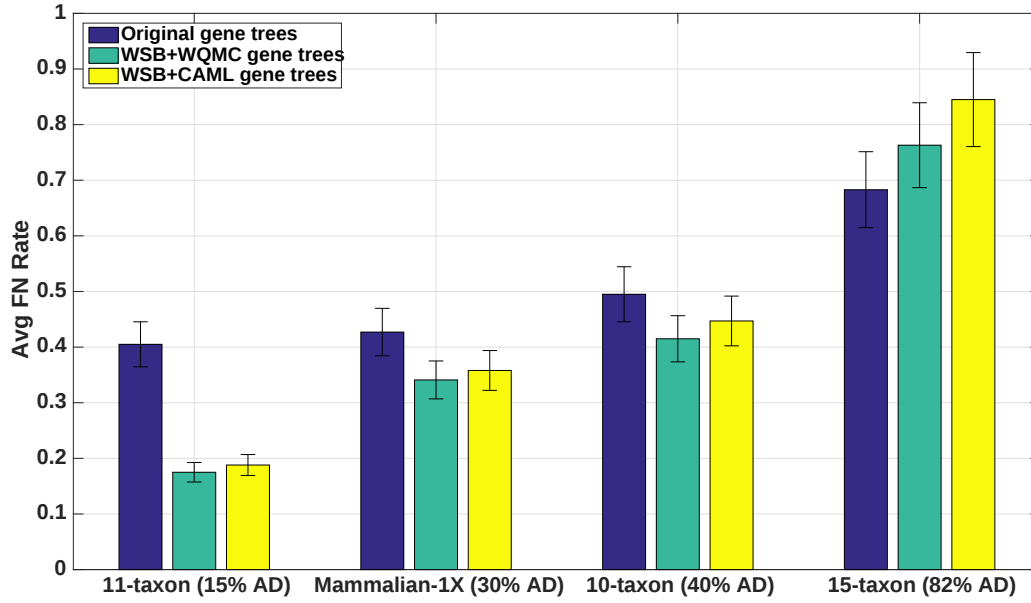


Figure 4.11: **Gene tree estimation error on 11-taxon, Mammalian-1X, 10-taxon and 15-taxon datasets for running WSB+WQMC and WSB+CAML.** We show results on the 11-taxon L ILS (100 genes and 500bp), Mammalian-1X (200 genes, 250bp), 10-taxon MH ILS (100 genes, 100bp), and 15-taxon H ILS (100 genes and 100bp) datasets were used. The level of ILS increases from left to right. WSB+WQMC was run using binning threshold 75% and confidence value 0.2. WSB+CAML was run using binning threshold 75%. The accuracy of WSB+WQMC and WSB+CAML gene trees were also compared to the accuracy of original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

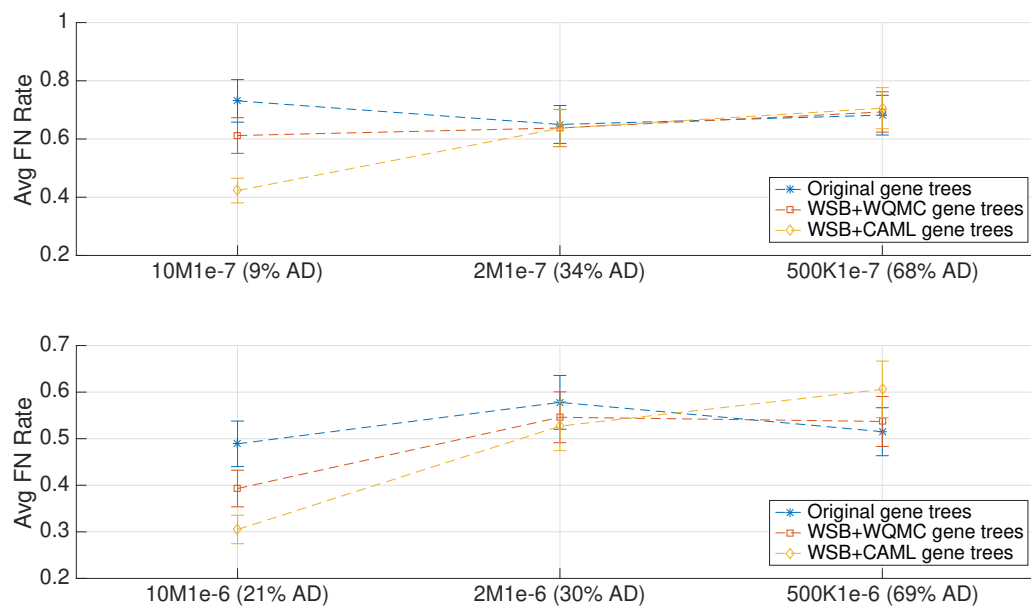


Figure 4.12: **Gene tree estimation error on SimPhy datasets having speciation rate $1e-7$ (top) and having speciation rate $1e-6$ (bottom) for running WSB+WQMC and WSB+CAML.** We show results on the 10M1e-7 (9% AD), 2M1e-7 (34% AD), 500KM1e-7 (68% AD), 10M1e-6 (21% AD), 2M1e-6 (30% AD), and 500K1e-6 (69% AD) datasets with 100 genes and 100bp alignment length. WSB+WQMC was run using binning threshold 95% and confidence value 0.2. WSB+CAML was run using binning threshold 95%. The accuracy of WSB+WQMC and WSB+CAML gene trees is also compared to the accuracy of original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

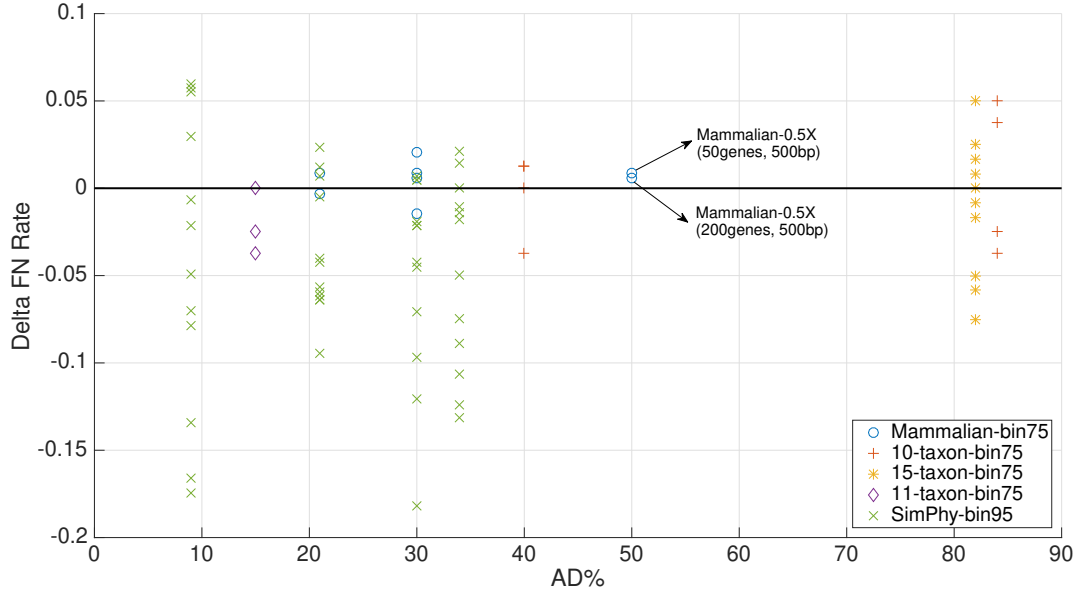


Figure 4.13: **ASTRAL2** species tree estimation error results on running **WSB+WQMC** on different model conditions of datasets shown in Tables 2.1 and 2.2. We show results on the Mammalian-0.5X (50% AD), Mammalian-1X (30% AD), Mammalian-2X (21% AD), SimPhy-10M1e-7 (9% AD), SimPhy-10M1e-6 (21% AD), SimPhy-2M1e-6 (30% AD), SimPhy-2M1e-7 (34% AD), 10-taxon MH ILS (40% AD), 10-taxon H ILS (84% AD), 15-taxon H ILS (82% AD), and 11-taxon L ILS (15% AD) datasets with different number of genes and alignment lengths, as shown in Tables 2.1 and 2.2. ILS level is varied on the x-axis with increasing ILS from left to right. Delta FN rate was computed by subtracting average FN rate of species tree on original gene trees from average FN rate of species tree on WSB+WQMC gene trees. Average FN rates were computed over 10 replicates for each dataset and model condition. Each point in the plot corresponds to delta FN rate on some model condition of the corresponding dataset. WSB+WQMC was run on these datasets with confidence value 0.2 and binning thresholds indicated in the legend.

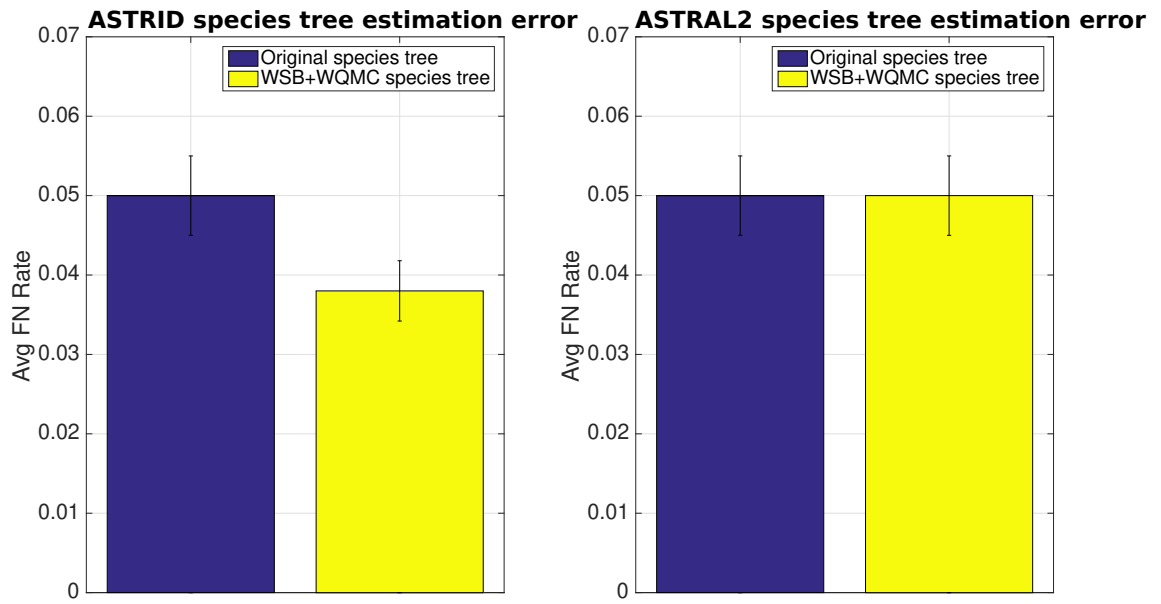


Figure 4.14: **Species tree estimation error using ASTRID (left) and ASTRAL2 (right) on 11-taxon dataset for running WSB+WQMC.** We show results on the 11-taxon L ILS (15% AD) dataset was used having 100 genes and 500bp alignment length. WSB+WQMC was run using binning threshold 75% and confidence value 0.2. The accuracy of WSB+WQMC gene trees and species tree is also compared to the accuracy of original gene trees and species tree. Average FN rate is shown with standard error bars over 10 replicates.

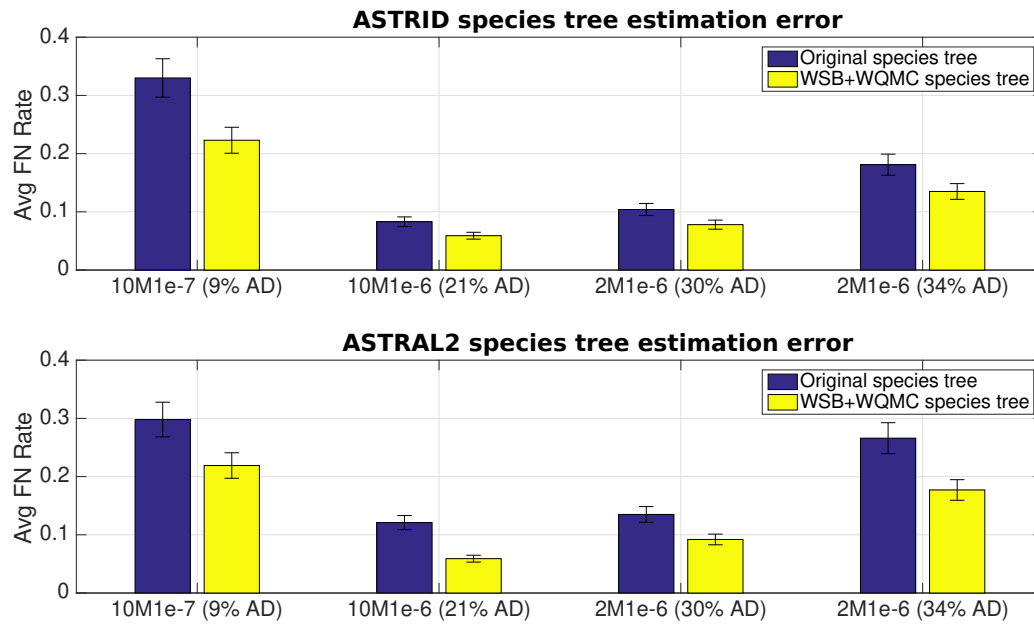


Figure 4.15: **Species tree estimation error using ASTRID (top) and ASTRAL2 (bottom) on SimPhy datasets for running WSB+WQMC.** We show results on the 10M1e-7 (9% AD), 10M1e-6 (21% AD), 2M1e-6 (30% AD), and 2M1e-7 (34% AD) datasets with 200 genes and 200bp alignment length. WSB+WQMC was run using binning threshold 95% and confidence value 0.2. The accuracy of species tree on WSB+WQMC gene trees is also compared to the accuracy of species tree on original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

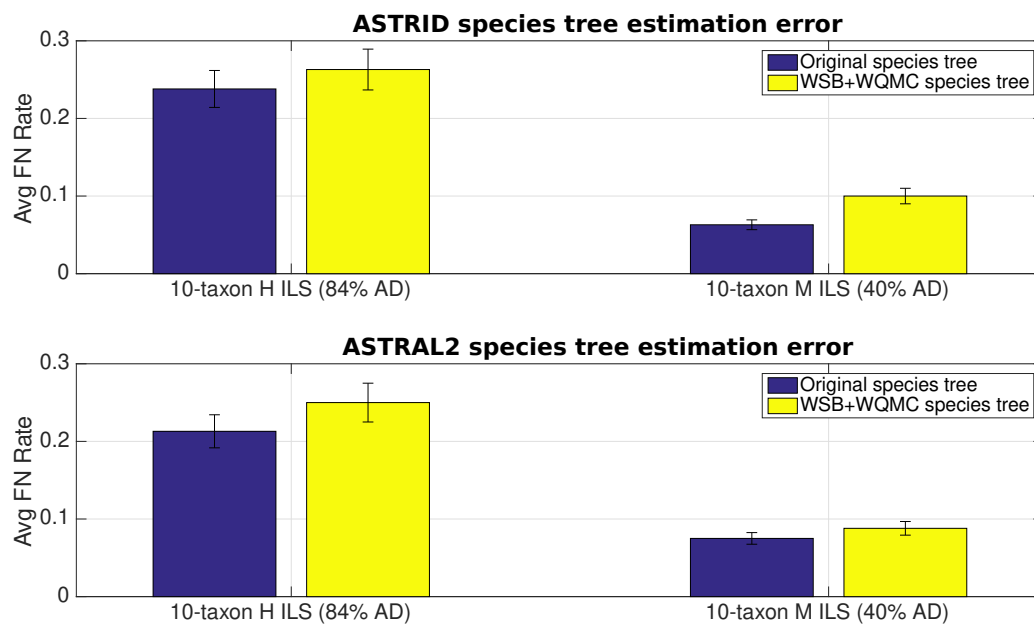


Figure 4.16: **Species tree estimation error using ASTRID (top) and ASTRAL2 (bottom) on 10-taxon datasets for running WSB+WQMC.** We show results on the 10-taxon MH ILS (40% AD) and 10-taxon H ILS (84% AD) with 200 genes and 100bp alignment length. WSB+WQMC was run using binning threshold 75% and confidence value 0.2. The accuracy of species tree on WSB+WQMC gene trees is also compared to the accuracy of species tree on original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

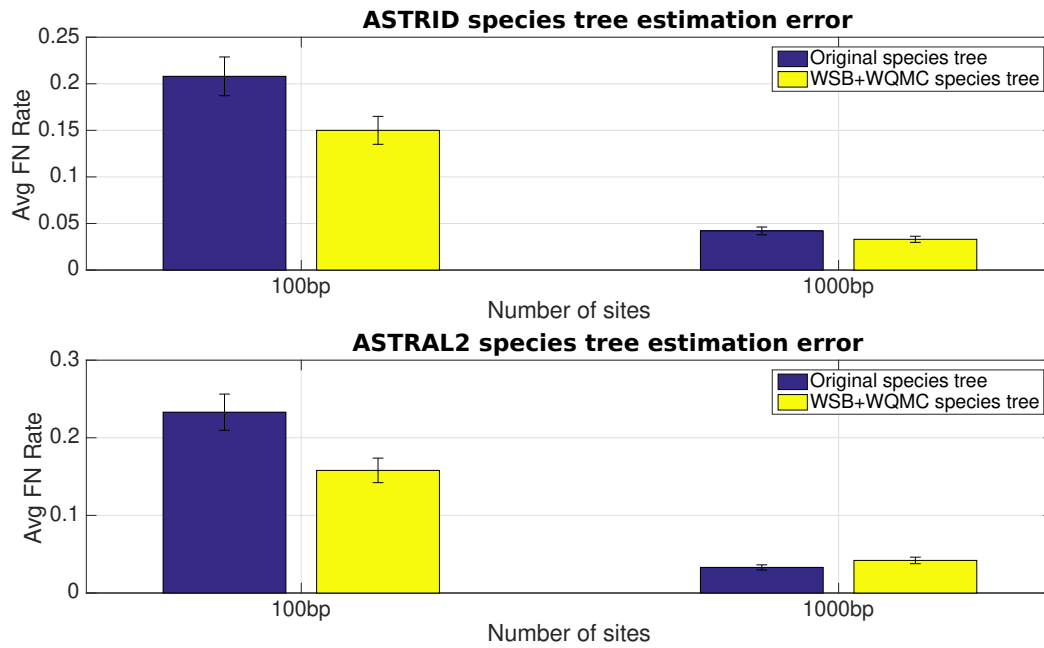


Figure 4.17: **Species tree estimation error using ASTRID (top) and ASTRAL2 (bottom) on 15-taxon H ILS dataset with two different alignment lengths for running WSB+WQMC.** We show results on the 15-taxon H ILS (82% AD) dataset was used having 1000 genes and two different alignment lengths (100bp and 1000bp). WSB+WQMC was run using binning threshold 75% and confidence value 0.2. The accuracy of species tree on WSB+WQMC gene trees is also compared to the accuracy of species tree on original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

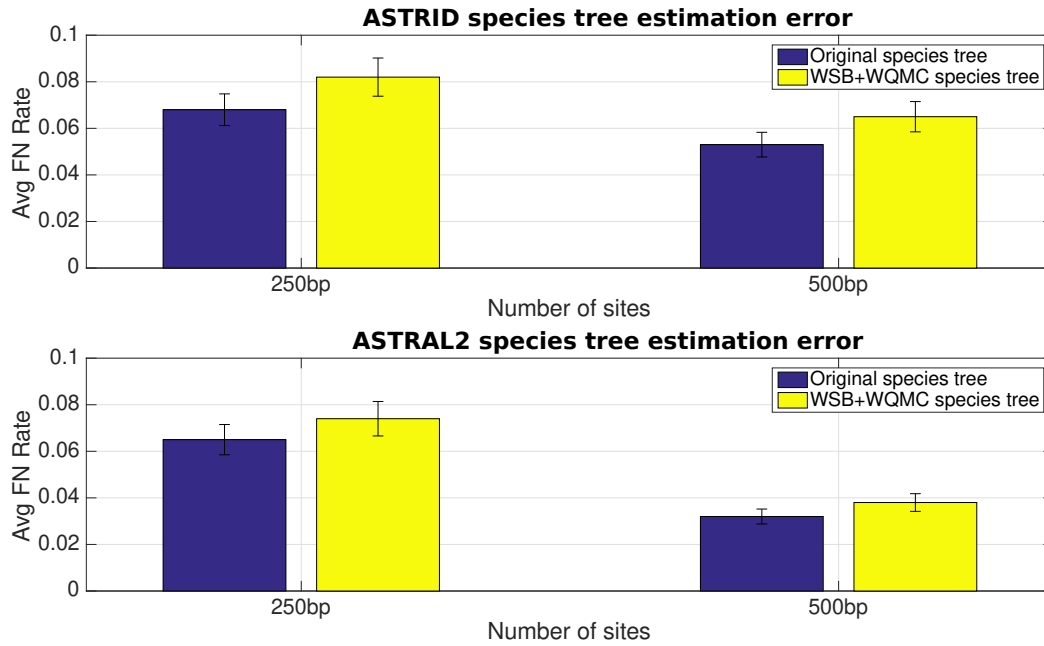


Figure 4.18: **Species tree estimation error using ASTRID (top) and ASTRAL2 (bottom) on the Mammalian-1X dataset with two different alignment lengths for running WSB+WQMC.** We show results on the Mammalian-1X (30% AD) dataset was used having 200 genes and two different alignment lengths (250bp and 500bp). WSB+WQMC was run using binning threshold 75% and confidence value 0.2. The accuracy of species tree on WSB+WQMC gene trees is also compared to the accuracy species tree on original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

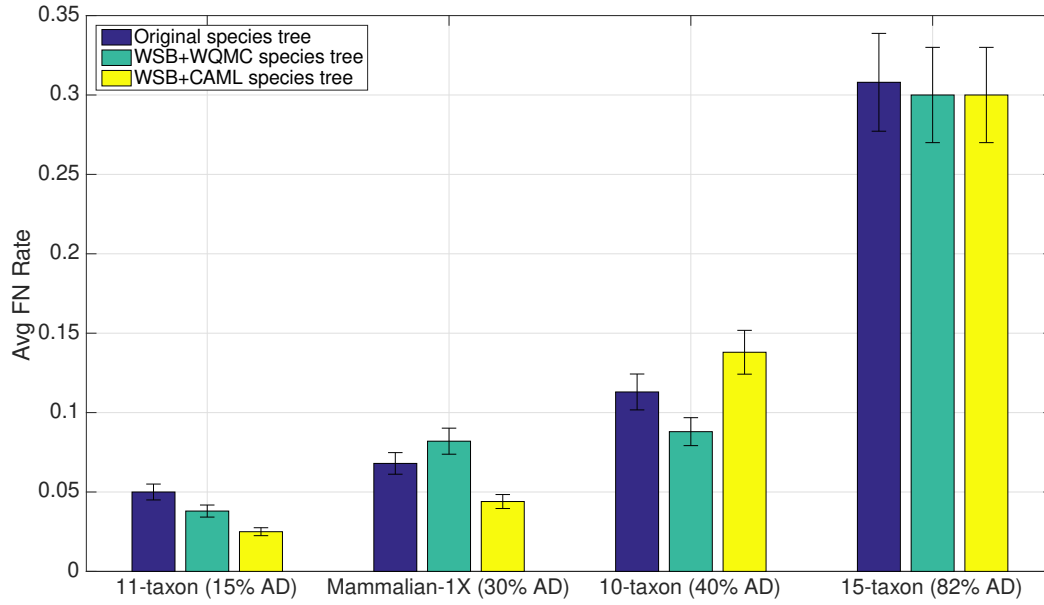


Figure 4.19: **ASTRID** species tree estimation error on 11-taxon, Mammalian-1X, 10-taxon and 15-taxon datasets for running WSB+WQMC and WSB+CAML. We show results on the 11-taxon L ILS (100 genes and 500bp), Mammalian-1X (200 genes, 250bp), 10-taxon MH ILS (100 genes, 100bp), and 15-taxon H ILS (100 genes and 100bp) datasets were used. The level of ILS increases from left to right. WSB+WQMC was run using binning threshold 75% and confidence value 0.2. WSB+CAML was run using binning threshold 75%. The accuracy of species tree on WSB+WQMC and WSB+CAML gene trees were also compared to the accuracy of species tree on original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

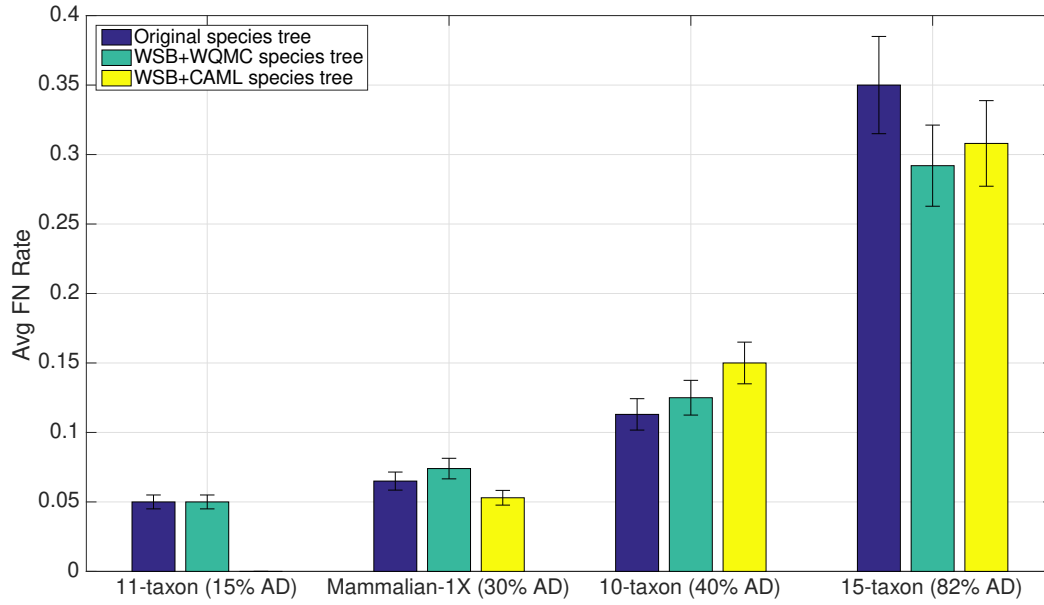


Figure 4.20: **ASTRAL2 species tree estimation error on 11-taxon, Mammalian-1X, 10-taxon, and 15-taxon datasets for running WSB+WQMC and WSB+CAML.** We show results on the 11-taxon L ILS (100 genes and 500bp), Mammalian-1X (200 genes, 250bp), 10-taxon MH ILS (100 genes, 100bp) and 15-taxon H ILS (100 genes and 100bp) datasets were used. The level of ILS increases from left to right. WSB+WQMC was run using binning threshold 75% and confidence value 0.2. WSB+CAML was run using binning threshold 75%. The accuracy of species tree on WSB+WQMC and WSB+CAML gene trees were also compared to the accuracy of species tree on original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

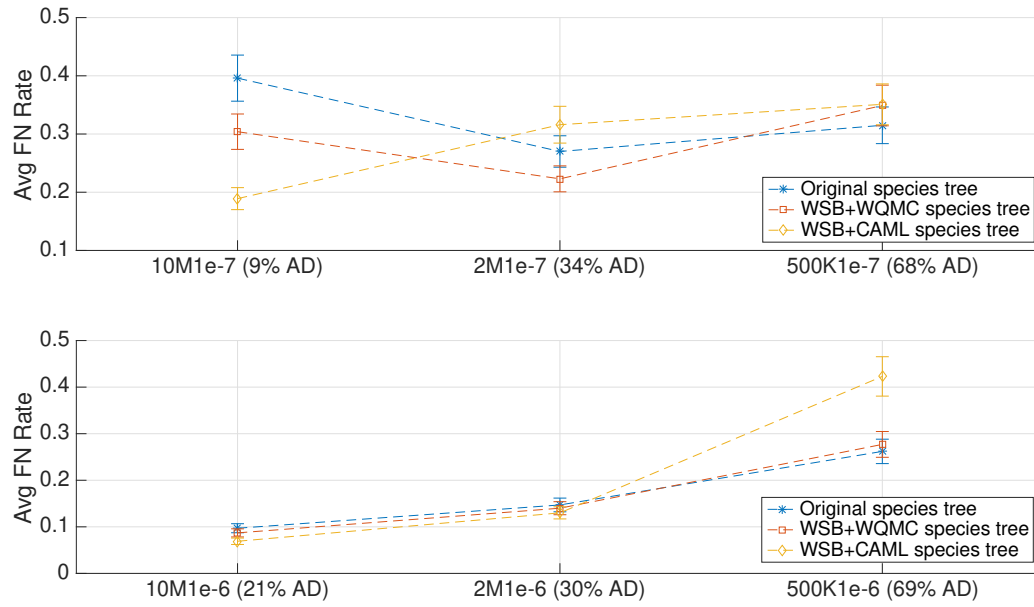


Figure 4.21: **ASTRID species tree estimation error on SimPhy datasets having speciation rate 1e-7 (top) and having speciation rate 1e-6 (bottom) for running WSB+WQMC and WSB+CAML.** We show results on the 10M1e-7 (9% AD), 2M1e-7 (34% AD), 500KM1e-7 (68% AD), 10M1e-6 (21% AD), 2M1e-6 (30% AD), and 500K1e-6 (69% AD) datasets with 100 genes and 100bp alignment length. WSB+WQMC was run using binning threshold 95% and confidence value 0.2. WSB+CAML was run using binning threshold 95%. The accuracy of species tree on WSB+WQMC and WSB+CAML gene trees is also compared to the accuracy of species tree on original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

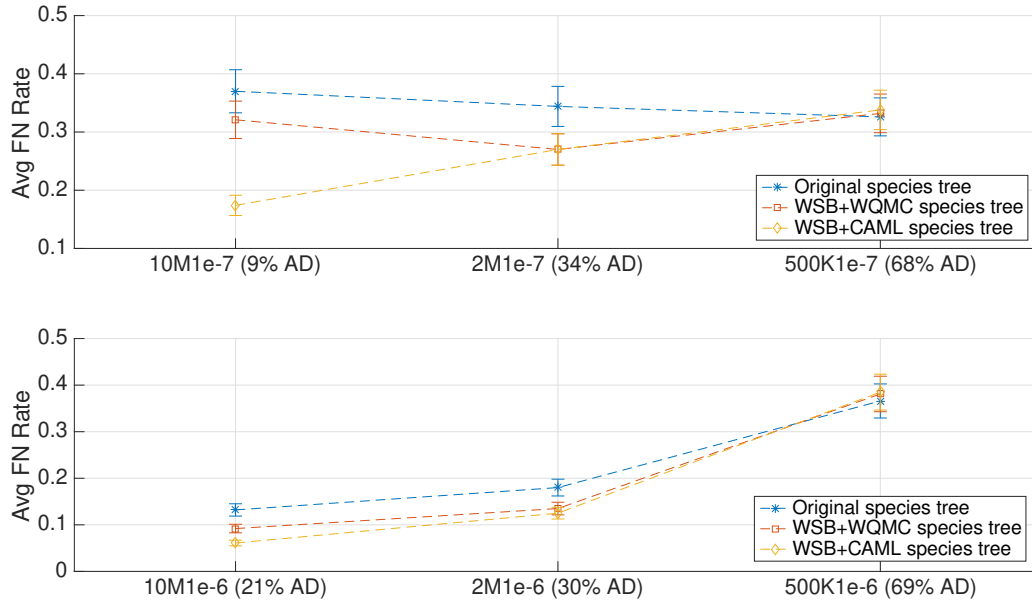


Figure 4.22: **ASTRAL2 species tree estimation error on SimPhy datasets having speciation rate 1e-7 (top) and having speciation rate 1e-6 (bottom) for running WSB+WQMC and WSB+CAML.** We show results on the 10M1e-7 (9% AD), 2M1e-7 (34% AD), 500KM1e-7 (68% AD), 10M1e-6 (21% AD), 2M1e-6 (30% AD), and 500K1e-6 (69% AD) datasets with 100 genes and 100bp alignment length. WSB+WQMC was run using binning threshold 95% and confidence value 0.2. WSB+CAML was run using binning threshold 95%. The accuracy of species tree on WSB+WQMC and WSB+CAML gene trees is also compared to the accuracy of species tree on original gene trees. Average FN rate is shown with standard error bars over 10 replicates.

Chapter 5

Discussion

In this thesis, the WSB+WQMC pipeline was proposed and evaluated on many simulated datasets using BestML analysis. The WSB+CAML pipeline proposed in [4] was also evaluated on those datasets and compared with the WSB+WQMC pipeline using BestML analysis. To the best of my knowledge, this is the first study to evaluate the effect of phylogenetic pipelines that use any variants of the statistical binning technique in the BestML paradigm. The prior studies of statistical binning [3] and weighted statistical binning [4] only evaluated these pipelines using MLBS, which may not be the best way to compute gene trees and species tree in certain model conditions. In [22], it was seen that BestML computed more accurate species trees than MLBS in model conditions having large enough number of genes. This clearly highlights the importance of the experimental study conducted in this paper, which provides performance results of these binning-based pipelines in BestML paradigm for the first time.

5.1 Summary of observations

We begin with a comparison of the WSB+WQMC pipeline to the basic unbinned pipeline. In general, it was found that WSB+WQMC improved the accuracy of gene trees (see Figure 4.5) in all model conditions of datasets having low ILS and in most of the model conditions of datasets having medium and moderately high levels of ILS. On the other hand, in high ILS conditions WSB+WQMC computed gene trees that were worse than original gene trees. WSB+WQMC also improved the accuracy of species tree estimation using both ASTRID and ASTRAL2 on most model conditions of datasets having low, medium, and moderately high ILS conditions (see Figures 4.13, 4.14, and 4.15), with a few exceptions. Similar to gene tree estimation, it was observed that on most model conditions of datasets having high levels of ILS, WSB+WQMC computed worse species trees than the species trees on original gene trees for both ASTRAL2 and ASTRID (see Figures 4.13 and 4.16).

A comparison of WSB+WQMC to WSB+CAML in terms of gene tree estimation error provides some

interesting differences. In high ILS datasets, WSB+WQMC was found to compute more accurate gene trees than WSB+CAML. In moderately high ILS conditions WSB+WQMC was found to be at least as good as WSB+CAML. In low and medium ILS datasets, there was no clear winner and the relative performance of WSB+CAML and WSB+WQMC depends on the dataset itself.

When comparing WSB+WQMC and WSB+CAML in terms of species tree estimation more interesting trends were observed. In high ILS datasets, WSB+WQMC was at least as accurate as WSB+CAML for computing species trees using both ASTRID and ASTRAL2. In moderately high ILS conditions, again WSB+WQMC was found to be at least as good as WSB+CAML, irrespective of the summary method. In low and medium ILS datasets WSB+CAML computed better species trees than WSB+WQMC using both ASTRAL2 and ASTRID.

5.2 Impact of ILS level on WSB+WQMC and WSB+CAML

For low enough ILS levels, both WSB+WQMC gene trees and WSB+CAML gene trees were more accurate than original gene trees, but as the ILS level increased the improvement decreased, until finally the re-estimated gene trees were less accurate. For species tree estimation, a similar relationship with ILS level was observed. This relationship of accuracy of binning based pipelines and ILS level was observed in previous studies as well. In [2], the naive binning pipeline was shown to improve species tree estimation in low ILS conditions. Similarly, both statistical binning and weighted statistical binning improved species tree estimation over the basic unbinned pipeline for low ILS conditions.

This relationship of ILS level and performance of WSB+WQMC and WSB+CAML is expected, as in low ILS conditions the bins computed by both pipelines will have genes having very similar evolutionary histories, and thus, using data from genes within the same bin will boost the phylogenetic signal and won't lead to any problematic analysis. However, when the level of ILS is high and sequences aren't very long, the binning step in these pipelines will group genes from different evolutionary histories into the same bin, leading to poor performance of both WSB+CAML and WSB+WQMC. This will happen mainly because many conflicting bipartitions between genes will not be detected in the binning step due to low bootstrap support values of edges in the initially estimated gene tree.

5.3 Impact of alignment length on WSB+WQMC

In this thesis it was observed that the magnitude of improvement in gene tree and species tree estimation when comparing the WSB+WQMC pipeline to the basic unbinned pipeline was greater for shorter alignment

lengths. This trend was also seen in [4], where it was seen that the difference between species tree estimation error by running the WSB+CAML pipeline and the basic unbinned pipeline decreased with increasing sequence length. This trend makes sense as when the alignment lengths are longer bootstrap support values of edges in the estimated gene trees increase, leading to lower bin sizes for the same binning threshold. Also, as the alignment becomes longer, the initially estimated gene tree improves and amount of improvement that can be achieved using binning pipelines is low.

5.4 Impact of binning threshold on WSB+WQMC

In this thesis, it was observed that binning 75% outperformed binning 100% in medium and high ILS conditions, with binning 100% being much worse and computing highly inaccurate gene trees and ASTRAL2 species tree for high ILS dataset. This is mainly because having a very high binning threshold (putting all genes into a single bin in this case) can put genes with very different evolutionary histories into the same bin. On the other hand, it was also seen that binning 100% performed better than binning 75% for low ILS datasets. This result can be attributed to the fact that lower binning threshold reduces bin sizes, which results in less data being used for each gene to compute a new gene tree. The limited amount of data used particularly hurts in low ILS conditions where all gene trees have very similar histories and having larger bin sizes can substantially improve performance of binning-based pipelines.

5.5 Relationship between gene tree and species tree estimation error

As expected, for most model conditions studied in this paper, improving gene trees using both WSB+WQMC and WSB+CAML leads to improvement in species tree estimation. However, it was observed that for some cases improving gene trees using WSB+WQMC and WSB+CAML didn't improve ASTRID and ASTRAL2 species trees when compared to the species trees on original gene trees. For example, the WSB+WQMC pipeline on the Mammalian-1X dataset computed better gene trees but worse species trees (see Figures 4.9 and 4.18). Moreover, WSB+CAML improved gene tree estimation on the 10-taxon dataset, but made species trees worse (see Figures 4.11, 4.19 and 4.20).

This was an unexpected result and conflicted with the common belief that reducing gene tree estimation error leads to reduction in species tree estimation error. The reason for this behavior remains unclear and needs further investigation in the future.

Another anomaly was observed when both WSB+WQMC and WSB+CAML improved ASTRAL2 and

ASTRID species trees for the 15-taxon H ILS dataset even after increasing gene tree estimation error (see Figures 4.11, 4.19 and 4.20). 15-taxon H ILS datasets have strict molecular clock and it is suspected that it can be leading to this strange result.

5.6 MLBS vs. BestML

The impact of ILS level and alignment length on the performance of binning-based pipelines were similar for both BestML and MLBS analyses. However, we found some important differences between BestML and MLBS that need to be discussed. In [4], it was found that the statistical binning and WSB+CAML pipelines decreased species tree accuracy compared to the basic unbinned pipeline only in a very few model conditions having ultra high levels of ILS (more than 80% AD) and small numbers of taxa. On the other hand, in BestML analyses we found that both WSB+WQMC and WSB+CAML were found to increase species tree estimation error compared to the basic unbinned pipeline more frequently and even in model conditions having 40% AD ILS level (see Figures 4.19 and 4.20). The benefit of using BestML over MLBS or vice versa is not clear and requires further research.

5.7 Using FastTree support values instead of bootstrap support values

The original gene trees on the SimPhy datasets were estimated without bootstrap support values. In order to run the WSB+WQMC and WSB+CAML pipelines on these datasets, FastTree local support values instead of bootstrap support values were used. FastTree support values tend to be much higher than bootstrap support values, which resulted in a very high bootstrap threshold value of 95% being used for the SimPhy datasets. Even though FastTree local support values of an edge gives some indication of the reliability of that edge, they are very different from bootstrap support values and are suspected to be of poor quality. Due to the use of FastTree local support values instead of bootstrap support values, there is a possibility that the trends seen on the SimPhy datasets may not hold if bootstrap support values are used in the future analysis.

5.8 Limitations of WSB+WQMC

Even though the WSB+WQMC pipeline decreased gene tree and species tree estimation error on many datasets studied in this paper, it has a few limitations. Similar to the WSB+CAML pipeline, it performed

poorly in high ILS conditions. Though it performed better than WSB+CAML in all high ILS datasets we examined, many times it was better to not use WSB+WQMC compared to the basic unbinned pipeline. This poor performance in high ILS conditions can be attributed to putting highly varied genes into a single bin, a common problem haunting binning-based pipelines.

Another drawback of WSB+WQMC is its need for the confidence value c and the binning threshold t . These values have a significant impact on the performance of WSB+WQMC, but we do not yet have a way of determining the optimal values that should be used in an experiment. In this study, the training phase was used to narrow down confidence values based on a subset of data and that value was used for the rest of the data. Even though we were able to obtain a good confidence value to be used for all the datasets in the study, the optimal confidence value can be different for different datasets.

Chapter 6

Summary

Gene tree estimation and species tree estimation are known to suffer when individual loci have low phylogenetic signal. In most phylogenetic studies, the limited amount of data causes the gene alignments to have low phylogenetic signal, which in turn leads to poorly estimated gene trees. Because species trees and gene trees can differ due to incomplete lineage sorting, the estimation of species trees requires multiple loci. In the presence of ILS, many summary methods like ASTRAL2, MP-EST, and ASTRID can be used to estimate the species tree by combining gene trees. However, this study, as well as others [19,20], suggests that species trees computed by summary methods using poorly estimated gene trees have low accuracy. Thus, an attempt to improve gene trees by enhancing phylogenetic signal helps in both gene tree and species tree estimation.

Three methods (naive binning, statistical binning, and weighted statistical binning) have been developed to improve gene tree and species tree estimation when individual gene alignments have low phylogenetic signal. The statistical binning pipeline proposed in [3] was found to improve species tree estimation using MLBS (multi-locus bootstrapping) analysis. However, in [23], the statistical binning pipeline was shown to be statistically inconsistent under the GTR+MSC model. The weighted statistical pipeline [4] improved on the statistical binning pipeline and was shown to be statistically consistent under the GTR+MSC model. In [4], the weighted statistical binning pipeline was evaluated on many datasets using MLBS analysis. When compared to the basic unbinned pipeline, the weighted statistical binning pipeline (referred as WSB+CAML in this thesis) was found to compute more accurate ASTRAL2 and MP-EST species trees except for a few datasets having extremely high ILS and small numbers of taxa.

Prior to this study, the weighted statistical binning pipeline was only tested using MLBS analysis, leaving it untested when it was run using BestML analysis. Although MLBS is more frequently used for species tree estimation, it may not be the most accurate way to compute species trees. The results in [22] showed that BestML computed more accurate species trees than MLBS with different species tree estimation methods on datasets having large enough number of genes. This indicates the importance of evaluating weighted

statistical binning and other phylogenetic pipelines using BestML analysis.

In this thesis, a novel phylogenetic pipeline named WSB+WQMC to improve gene tree and species tree estimation in the presence of low phylogenetic signal is proposed. The WSB+WQMC pipeline shares several design features with the WSB+CAML pipeline, but has some other desirable properties. The WSB+WQMC pipeline has good theoretical guarantees and is shown to be statistically consistent under the GTR+MSC model when a slightly different version of WQMC [23] is used in the pipeline.

An experimental study was conducted in this thesis to evaluate the WSB+WQMC pipeline. It also compared WSB+WQMC to the WSB+CAML pipeline in terms of gene tree and species tree estimation on various datasets. This study was the first to evaluate the weighted statistical binning pipeline using BestML analysis and found that both WSB+WQMC and WSB+CAML followed trends seen in MLBS analysis with some important differences.

When compared to the initially estimated gene trees, both WSB+WQMC and WSB+CAML pipelines were found to help in improving gene tree estimation in low ILS conditions. However, the level of ILS had a big effect on the performance of both pipelines, and it was found that both WSB+WQMC and WSB+CAML can make gene trees worse on high ILS datasets.

Furthermore, when compared to ASTRID and ASTRAL2 species trees on original gene trees, both WSB+WQMC and WSB+CAML pipelines again improved species tree estimation on most of the datasets studied. But, on a few datasets, both pipelines ended up reducing the accuracy of species trees. This result was unexpected and refutes the common belief that reducing gene tree estimation error always reduces species tree estimation error. Therefore, care must be taken when using both WSB+WQMC and WSB+CAML pipelines as they may not always turn out to be beneficial.

When comparing WSB+WQMC and WSB+CAML in terms of gene tree and species tree estimation, no clear winner was found. The relative performance of WSB+WQMC and WSB+CAML was dependent on the dataset and the level of ILS. In low and medium levels of ILS, it was found that WSB+CAML computed better species trees than WSB+WQMC, but there was no clear winner for gene tree estimation. On the other hand, in moderately high and high levels of ILS, WSB+WQMC was at least as good as WSB+CAML for both gene tree and species tree estimation.

This study clearly suggests the potential benefits of using both WSB+CAML and WSB+WQMC pipelines for gene tree and species tree estimation. However, there are a few cases where both these pipelines can rather hurt in estimating gene trees and species tree. This, warrants the need for refinements in both WSB+WQMC and WSB+CAML pipelines along with research in the direction of using binning-based pipelines to boost phylogenetic signal.

This thesis highlights various methods to improve gene tree and species tree estimation in the presence of

low phylogenetic signal. However, the decision to use binning-based pipelines rather than the basic unbinned pipeline is not straightforward and can have consequences. Moreover, the choice of using MLBS or BestML for running these pipelines is also not clear and requires further research.

One major challenge in this study was evaluating the performance of WSB+WQMC and WSB+CAML pipelines on the SimPhy datasets. The original gene trees for these datasets did not have bootstrap support values on edges and FastTree support values were used instead in the analysis. FastTree support values are suspected to be a lower quality estimator for edge reliability than bootstrap support values. This casts some doubts on the results obtained on the SimPhy datasets and demonstrates the limit of using these datasets in this thesis.

The WSB+WQMC pipeline is not perfect and has scope for improvements. It requires the confidence value and the binning threshold value to be set. Research in the direction of automatically inferring these parameters may substantially improve the accuracy of the pipeline by being able to pick optimal values for these parameters based on the data. The WSB+WQMC pipeline used WQMC as the heuristic to estimate gene trees from the set of weighted quartets and it would be interesting to explore the performance of the pipeline using other heuristics. The WSB+WQMC pipeline performed poorly in most high ILS datasets; thus, WSB+WQMC needs to be further improved in the future to avoid decreasing accuracy of gene trees in high ILS conditions.

This study only investigated the performance of WSB+WQMC on simulated datasets. The performance of the WSB+WQMC pipeline on biological datasets needs to be evaluated in the future. It was seen that binning threshold value has an impact on binning pipelines and research aiming at determining good threshold values would be beneficial. One question that isn't addressed in this study is whether it is better to use BestML or MLBS analysis for binning-based pipelines. Evaluating the WSB+WQMC and WSB+CAML pipelines using MLBS analysis can help answer this question. Finally, since concatenation is known to be computationally expensive, it is conjectured that WSB+WQMC is faster and more scalable than WSB+CAML. Analysis of running time and memory usage of the WSB+CAML and WSB+WQMC pipelines can help confirm our suspicion and needs to be done in the future.

Bibliography

- [1] Jonathan A Eisen. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, 8(3):163–167, 1998.
- [2] Md Shamsuzzoha Bayzid and Tandy Warnow. Naive binning improves phylogenomic analyses. *Bioinformatics*, 29(18):2277–2284, 2013.
- [3] Siavash Mirarab, Md Shamsuzzoha Bayzid, Bastien Boussau, and Tandy Warnow. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, 346(6215):1250463, 2014.
- [4] Md Shamsuzzoha Bayzid, Siavash Mirarab, Bastien Boussau, and Tandy Warnow. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. *PloS One*, 10(6):e0129183, 2015.
- [5] Wayne P Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 1997.
- [6] James H Degnan and Noah A Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6):332–340, 2009.
- [7] James H Degnan and Noah A Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS Genet*, 2(5):e68, 2006.
- [8] Scott V Edwards. Is a new and general theory of molecular systematics emerging? *Evolution*, 63(1):1–19, 2009.
- [9] Sebastien Roch and Mike Steel. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100:56–62, 2015.
- [10] James H Degnan, Michael DeGiorgio, David Bryant, and Noah A Rosenberg. Properties of consensus methods for inferring species trees from gene trees. *Systematic Biology*, page syp008, 2009.
- [11] Laura Salter Kubatko and James H Degnan. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1):17–24, 2007.
- [12] Liang Liu. Best: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, 24(21):2542–2543, 2008.
- [13] Alexei J Drummond and Andrew Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1):214, 2007.
- [14] Siavash Mirarab and Tandy Warnow. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12):i44–i52, 2015.
- [15] Pranjal Vachaspati and Tandy Warnow. ASTRID: Accurate species trees from internode distances. *BMC Genomics*, 16(Suppl 10):S3, 2015.
- [16] Liang Liu, Lili Yu, and Scott V Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1):302, 2010.
- [17] John Frank Charles Kingman. The coalescent. *Stochastic Processes and Their Applications*, 13(3):235–248, 1982.

- [18] Liang Liu and Dennis K Pearl. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology*, 56(3):504–514, 2007.
- [19] Siavash Mirarab, Rezwana Reaz, Md S Bayzid, Théo Zimmermann, M Shel Swenson, and Tandy Warnow. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, 2014.
- [20] Adam D Leaché and Bruce Rannala. The accuracy of species tree estimation under simulation: a comparison of methods. *Systematic Biology*, page syq073, 2010.
- [21] Simon Tavaré. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17:57–86, 1986.
- [22] Siavash Mirarab, Md Shamsuzzoha Bayzid, and Tandy Warnow. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, page syu063, 2014.
- [23] Eliran Avni, Reuven Cohen, and Sagi Snir. Weighted quartets phylogenetics. *Systematic Biology*, 64(2):233–242, 2015.
- [24] Mark A Ragan. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution*, 1(1):53–58, 1992.
- [25] Nam Nguyen, Siavash Mirarab, and Tandy Warnow. MRL and SuperFine+ MRL: new supertree methods. *Algorithms for Molecular Biology*, 7(1):1, 2012.
- [26] Tandy Warnow. *Computational Phylogenetics*. Draft of the book available online at <http://tandy.cs.illinois.edu/textbook.pdf>.
- [27] Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.
- [28] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2—approximately maximum-likelihood trees for large alignments. *PloS One*, 5(3):e9490, 2010.
- [29] Diego Mallo, Leonardo De Oliveira Martins, and David Posada. Simphy: Phylogenomic simulation of gene, locus, and species trees. *Systematic Biology*, page syv082, 2015.
- [30] Ziheng Yang. Mcoale. 2015. Software available online at <http://abacus.gene.ucl.ac.uk/software/MCMCcoal.html>.
- [31] Yujin Chung and Cécile Ané. Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. *Systematic Biology*, page syr003, 2011.