Title

New methods for inferring, assessing, and using phylogenetic trees from genomic and microbiome data

Author
Sayyari, Erfan

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

New methods for inferring, assessing, and using phylogenetic trees from genomic and microbiome data

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Signal and Image Processing)

by

Erfan Sayyari

Committee in charge:

Professor Siavash Mirarab, Chair
Professor Vineet Bafna
Professor Alon Orlitsky
Professor Greg Rouse
Professor Behrouz Touri

2019

The dissertation of Erfan Sayyari is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

DEDICATION

I dedicate my dissertation to my beloved mother and father.

EPIGRAPH

Nothing in Biology Makes Sense Except in the Light of Evolution.

—Theodosius Dobzhansky

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my very great appreciation to my collaborators and mentors who helped me throughout my Ph.D. I would like to extend my sincere thanks to my labmates particularly Maryam Rabiee, Chao Zhang, Metin Balaban, and Uyen Mai whom I enjoyed collaborating with. I would like to express my deepest appreciation to my committee. I am also grateful for the computational support from the San Diego Supercomputer Center (SDSC) through XSEDE allocations, which is supported by the NSF. I must thank the financial support I received throughout my Ph.D. from NSF, the Center for Microbiome Innovation (CMI), and IBM Research AI through the AI Horizons Network.

The completion of my dissertation would not have been possible without the support and nurturing of my advisor Dr. Siavash Mirarab. As his first graduate student, I had the chance to closely work with and learn much from him and achieved accomplishments that I would have never imagined before. He is a great advisor and the most caring and patient mentor I could have asked for. I owe him much of my current and future achievements.

Finally, I would thank my siblings and parents for their emotional support throughout my Ph.D. I wouldn't imagine achieving any of my success without my parents' prayers and teachings. My mother showed me how one can be generous, kind, and thankful, even in the most difficult situations. I also have always benefited from the insightful guidance of my father to achieve wholesome living by considering important aspects of life other than my career.

Chapter 2, in full, contains material from SAYYARI, E., AND MIRARAB, S. Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction. BMC Genomics 17, S10 (2016), 101–113. I was the primary investigator and author of this paper.

Chapter 3, in full, contains material from SAYYARI, E., AND MIRARAB, S. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. Molecu-

lar Biology and Evolution 33, 7 (2016), 1654–1668. I was the primary investigator and author of this paper. I also want to extend my sincere thanks to Prof. Tandy Warnow and Prof. Jim Leebens-mack for their consistent support and helpful discussions to shape the content of this chapter.

Chapter 4, in full, contains material from SAYYARI, E., AND MIRARAB, S. Testing for polytomies in phylogenetic species trees using quartet frequencies. Genes 9, 3 (8 2018), 132. I was the primary investigator and author of this paper.

Chapter 5, in full, contains material from SAYYARI, E., WHITFIELD, J. B., AND MIRARAB, S. Fragmentary Gene Sequences Negatively Impact Gene Tree and Species Tree Reconstruction. Molecular Biology and Evolution 34, 12 (2017), 3279–3291. I was the primary investigator and author of this paper. I must thank Prof. James B. Whitfield, as the co-author of this publication from whom I learned a lot during our discussions and conversations. I also thank Prof. Bernhard Misof, Prof. Karl Kjer, and Dr. Karen Meusemann for providing the insect dataset and additional explanations of how it was constructed.

Chapter 6, in full, contains material from SAYYARI, E., WHITFIELD, J. B., AND MIRARAB, S. DiscoVista: Interpretable visualizations of gene tree discordance. Molecular Phylogenetics and Evolution 122 (2018), 110–115. I was the primary investigator and author of this paper.

Chapter 7, in full, contains material from SAYYARI, E., KAWAS, B., AND MIRARAB, S., TADA: Phylogenetic Augmentation of Microbiome Samples Enhances Phenotype Classification. Bioinformatics (ISMB Species Issue) that is accepted at ISMB 2019 and will appear in Bioinformatics journal. I was the primary investigator and author of this paper. I'm extremely grateful to Prof. Rob Knight, Dr. Sandrine Javorschi-Miller, and Dr. Ho-Cheol Kim, for their support and comments in developing the materials of this chapter. I should sincerely thank Dr. Austin Swafford and Dr. Ban Kawas, who mentored me during my collaboration with CMI and IBM Research AI through the AI Horizons Network. I enjoyed having insightful discussions and conversations with them that deeply broadened my knowledge of ongoing research in the microbiome and machine learning communities.

VITA

2008-2013	B.Sc. in Electrical Engineering, Sharif University of Technology, Tehran, Iran
2013-2016	M.Sc. in Electrical Engineering (Signal and Image Processing), University of California, San Diego
2016-2019	Ph.D. in Electrical Engineering (Signal and Image Processing), University of California, San Diego

PUBLICATIONS

SAYYARI, E., KAWAS, B., AND MIRARAB, S., TADA: Phylogenetic Augmentation of Microbiome Samples Enhances Phenotype Classification. Bioinformatics (ISMB Species Issue). This manuscript is accepted at ISMB 2019 and is to appear in Bioinformatics journal.

RABIEE, M., SAYYARI, E., AND MIRARAB, S. Multi-allele species reconstruction using ASTRAL. Molecular Phylogenetics and Evolution 130 (2019), 286–296.

ZHANG, C., RABIEE, M., SAYYARI, E., AND MIRARAB, S. ASTRAL-III: poly- nomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19, S6 (2018), 153.

SAYYARI, E., AND MIRARAB, S. Testing for polytomies in phylogenetic species trees using quartet frequencies. Genes 9, 3 (8 2018), 132.

SAYYARI, E., WHITFIELD, J. B., AND MIRARAB, S. DiscoVista: Interpretable visualizations of gene tree discordance. Molecular Phylogenetics and Evolution 122 (2018), 110–115

SAYYARI, E., WHITFIELD, J. B., AND MIRARAB, S. Fragmentary Gene Sequences Negatively Impact Gene Tree and Species Tree Reconstruction. Molecular Biology and Evolution 34, 12 (2017), 3279–3291.

ZHANG, C., SAYYARI, E., AND MIRARAB, S. ASTRAL-III: Increased Scalability and Impacts of Contracting Low Support Branches. In Lecture Notes in Computer Science, J. Meidanis and L. Nakhleh, Eds., vol. 10562 LNBI. Springer International Publishing, Cham, 2017, pp. 53–75.

MAI,U.,SAYYARI,E.,ANDMIRARAB,S.Minimumvariancerootingofphylogenetic trees and implications for species tree reconstruction. PLOS ONE 12, 8 (2017), e0182238.

SAYYARI, E., AND MIRARAB, S. Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction. BMC Genomics 17, S10 (2016), 101–113.

SAYYARI, E., AND MIRARAB, S. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. Molecular Biology and Evolution 33, 7 (2016), 1654–1668.

WEIBEL, N., HWANG, S., RICK, S., SAYYARI, E., LENZEN, D., AND HOLLAN, J. Hands That Speak: An Integrated Approach to Studying Complex Human Commu- nicative Body Movements. In 2016 49th Hawaii International Conference on System Sciences (HICSS) (2016), pp. 610–619.

SAYYARI, E., FARZI, M., ESTAKHROOEIEH, R. R., SAMIEE, F., AND SHAMSOL- LAHI, M. B. Migraine analysis through EEG signals with classification approach. In 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA) (2012), pp. 859–863.

ABSTRACT OF THE DISSERTATION

New methods for inferring, assessing, and using phylogenetic trees from genomic and microbiome data

by

Erfan Sayyari

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California San Diego, 2019

Professor Siavash Mirarab, Chair

Phylogenies are trees showing the evolutionary relationship among species, and reconstructing phylogenies using molecular data can be framed as an optimization problem. Recent advances in DNA sequencing have resulted in extensive application of phylogenetic inference to (meta)genomic data. However, the scale and the complexity of the data has presented researchers with new algorithmic and statistical challenges, in particular, difficulties in noise reduction and statistical support estimation. This dissertation addresses these challenges.

A significant challenge in using genomic data for phylogenetics (phylogenomics) is inconsistencies between evolutionary histories across different parts of the genome. Thus,

phylogenomics methods need to consider these inconsistencies. One scalable solution is using summary methods, where a tree is first inferred for each gene, and then gene trees are summarized to build the species tree. Chapter 2 of this dissertation is dedicated to presenting a scalable and accurate summary method called DISTIQUE for reconstructing species trees from gene trees.

A major challenge in phylogenomics is the interpretation of inferred phylogenies, especially in the presence of noise and gene-tree inconsistencies. Biologists rely on measures of statistical support for interpreting branches of the phylogeny. Chapter 3 introduces a highly scalable and reliable Bayesian measure of support, localPP, and Chapter 4 introduces a frequentist version of localPP for performing hypothesis testing.

When using any summary method, the quality of the inferred species tree is highly impacted by the quality of gene phylogenies. In Chapter 5, we identify one factor that reduces the gene tree accuracy (gene fragmentation) and introduce a filtering strategy that effectively reduces error in gene trees and species trees. Further, Chapter 6 introduces a visualization framework, DiscoVista, to assist biologists in interpreting potentially discordant phylogenetic results.

The final chapter focuses on the use of phylogenies in microbiome studies, where the goal is analyzing genetic material from environmental samples and to infer associations of genotype to phenotypical properties of samples. A main challenge in microbiome analyses is the huge variability across samples and small sample sizes. Chapter 7 introduces TADA, a new phylogeny-based method of data augmentation that improves the accuracy of classification methods applied to microbiome data.

Chapter 1

Introduction

Phylogenies are typically trees showing evolutionary relationship between species. They are crucial for better understanding of the life diversity on earth [50]. More generally, phylogenies help us to explore how sequences, genomes, and species evolve [94]. The branching structure of phylogenies can represent the speciation events (i.e., the species tree), or the relationship between different genes (i.e., the gene trees). Also, the branch length shows the time between two nodes in evolutionary time. The leaves of the phylogeny can show the extant species (or genes), internal nodes present ancestral species (or genes), and the root of the phylogeny shows the common ancestor of all extant species.

We can represent species with character sequences, where these characters capture the characteristics of the species. Reconstructing the phylogenies is typically framed as reconstructing a tree using statistical models that best describes observed sequences for extant species [94, 68, 135]. Historically, morphology (e.g., color, shape, pattern, size) were used to reconstruct phylogenies. However, evolution can happen through accumulation of change in the genome of species, and thus we could use molecular data (e.g., DNA or protein sequences) to reconstruct phylogenies [94].

With advances in sequencing technology as well as advances in computational methods (e.g., development of ML methods for phylogenetic inferences), single gene sequences are used to infer the phylogenies since mid 20th century [67]. More recently and after introduction of technologies such as Next Generation Sequencing [157, 30]), using genomic data to reconstruct the phylogenies is becoming routine. Using whole genomes to infer phylogenies provides us tremendous amount of data and the prospect that we could solve phylogenetic problems with less uncertainties [76]. However, using whole genome (i.e., multiple loci) to infer phylogenies had introduced new challenges and complexities [106, 183]. The biological reasons behind these complexities lie with the inconsistencies prevalent between the evolutionary history of different parts of the genome (i.e. gene trees), and the phylogeny showing the speciation events (i.e., the species tree). To address inconsistencies prevalent between gene trees and species tree, new models of gene tree evolution are introduced [60, 74, 56].

Inferring phylogenies from sequence data typically requires multiple steps. Due to mutations (such as insertion and deletion) and for a particular gene, different species can have different sequence length. Therefore, a main step in many phylogenetic analyses is to align different species sequence belong to a particular gene so that they all have the same length using Multiple Sequence Alignment (MSA) algorithms [205, 137, 136]. The aligned sequences can be used to infer phylogenies. In this dissertation, I mostly focus on the pipelines of reconstructing species trees from genomic data. Previously, different methods of reconstructing species tree from genomic data are developed. One of these methods is called *concatenation*. In this method all genes simply concatenated together to form a *supermatrix*, and then a statistical phylogeny reconstruction method is used to infer the tree.

An alternative approach which directly considers gene tree discordance is the so-called two-step method (also called the summary method). In summary methods, i) for each gene a tree is constructed independently (e.g., using maximum likelihood methods [232, 188, 86]), and then ii) all gene trees are summarized to construct a tree for the entire genome (considering gene trees discordance). Since summary methods use gene trees to reconstruct the species tree, the quality of inferred species trees are largely affected by the quality of gene trees and missing data as I will elaborate in Chapter 5.

Finally, in biological analyses, we don't have access to the true phylogeny. Moreover, my inferences are based on limited noisy sequences. Thus, assessing the quality of reconstructed phylogenies is a major step in phylogenetic analyses.

In this dissertation, I contributed to different parts of phylogeny reconstruction pipeline, by introducing new methods of filtering, inferring, assessing and visualization phylogenies and phylogenomic data. I show that my methods can effectively improve the quality of phylogenies, enrich our understanding of the inconsistencies in the genomic data, and enables us to compute a reliable statistical support for datasets with large number of genes and species.

One of the major sources of discordance between gene trees and species tree is Incomplete Lineage Sorting (ILS, will be elaborated in Chapter 2). To address ILS, different summary methods of reconstruction species trees are introduced [277, 255, 142]. However, due to the scalability and accuracy issues introducing new methods is still needed. One principal way of inferring species tree is to first compute distances between species and then build a tree that best represents these estimated distances using distance-based methods like neighbor-joining (NJ) or UPGMA [206, 224]. In Chapter 2, I introduce a new summary method of estimating species tree from gene trees called DISTIQUE. I first show that by choosing two random species as anchors, I can calculate relative phylogenetic distances between all other pairs of species using gene trees (by considering unrooted trees of size four, i.e., *quartets*). Then, using distance-based methods like NJ, I can infer the species tree from computed distances.

To reduce the estimation bias and variance for the anchored distances, I either average the relative distances for randomly chosen anchors, or I build separate trees for different pairs of anchors and combine trees to get the final species tree. This method is guaranteed to produce the true species tree with high probability if ILS is the only source of discordance (i.e., a statistically consistent method under ILS). In my simulation and biological analyses, I compared DISTIQUE with the state-of-the-art summary methods including ASTRID [255] and ATRAL-II [165]. I showed that DISTIQUE is as accurate as (or sometimes better than) ASTRID but slightly less accurate than ASTRAL-II. In terms of running time, DISTIQUE is faster than ASTRAL-II but less scalable than ASTRID.

Summary methods first infer gene trees and then use them to reconstruct the species tree. Accordingly, the quality of summary methods is highly affected by the quality of gene trees. However, the bioinformatics pipelines used in the whole-genome and transcriptomic data preparation (i.e., transcribed part of the genome), can inject errors and missing data into the datasets. These errors can negatively affect the quality of reconstructed phylogeny [130, 185, 97]. For summary methods, two types of missing data exit: i) missing taxon, and ii) fragmentary data [97]. When for a gene, a species is entirely missing, we call it a missing taxon, and when a large portion of data is missing (i.e. fragmentation) we call it fragmentary data. In Chapter 5, I further analyzed the effects of fragmentary data on the quality of gene trees and eventually on the quality of species tree estimation. Then I introduced an effective filtering strategy to deal with fragmentation. I used biological and simulation analyses to benchmark the parameters and investigate the effectiveness of my filtering strategy. In particular, in simulations I showed that applying more filtering improves the quality of gene trees at the expense of introducing more missing taxa. Hence, there is a tradeoff between quality of species tree and applying more strict filtering.

Evaluating statistical support for reconstructed phylogenies is the essence of many phylogenomic analyses. Different methods of calculating support like multi-locus bootstrapping (MLBS, will be discuss in Chapter 3) [214] and Bayesian posterior probabilities [192, 124, 151] are widely used in summary methods. However, shortcomings of these methods such as lack of interpretability, scalability, and reliability motivate us to introduce novel measures of assessing phylogenies [159, 13, 221]. More particularly, calculating

Bayesian posterior probabilities requires MCMC sampling from the distribution of gene trees and species trees parameters and topology, and is computationally challenging. On the other hand, using MLBS, the support can be easily underestimated or overestimated [159, 221]. These biases are due to the higher level of gene tree discordance and errors derived by the lack of phylogenetic signal in bootstrapped gene alignments.

Under the MSC model of ILS (will be discussed in Chapter 2), and considering only four species (i.e., quartets with only three possible unrooted topology), the most frequent gene tree is topologically congruent with the species tree. Moreover, based on the MSC model the other alternative topologies are equiprobable with probabilities depending on the branch length separating species [6]. In Chapter 3, I analytically computed the posterior probability of different species tree topologies around a single branch given the inferred gene trees as input. Having more than four species, computing the posterior probability becomes computationally challenging [6, 151], and so I compute an approximate solution by mapping the problem to the problem of having four species. This measure of support is called local posterior probability (*localPP*). In simulations and biological analyses, I showed that localPP is highly reliable, and is more accurate than MLBS. Moreover, the algorithm I introduced to calculate localPP is highly scalable and can handle datasets with thousands of species and genes.

Typically, phylogenies are binary, meaning that all internal nodes of the tree are of degree three. In this case, we call the phylogeny a bifurcating tree. On the other hand, we could have trees that they have at least a node with degree greater than three (i.e., a *polytomy*). This could happen for two reasons, first we didn't have enough evidence and information to resolve the topology (i.e., *soft polytomy*). Alternatively, a speciation event can lead to more than two offspring, and thus produce a *hard polytomy*. In Chapter 4, I introduce a statistical hypothesis test for detecting multifurcating events. The test null hypothesis is that the branch length is zero and the branch is *multifurcating* (i.e., forms a polytomy). This test is a frequentist version of localPP, and incorporates the same mapping introduced in Chapter 3. In simulations and biological analyses, I show the effectiveness of my method.

With the amount of complexity and incongruence prevalent in the phylogenomic data, visualizing the results and making comparisons between different studies is crucial. Despite the availability of tools for comparing and visualizing phylogenomic and phylogenetic analyses, the interpretation of their visualizations and measures remains challenging. In Chapter 6 I introduce *DiscoVista* to create easily interpretable visualizations to compare different trees (from various studies or using different methods) and to show the amount of discordance and biases available in phylogenomic data. Using biological datasets, I show different modules and visualizations of DicoVista.

Phylogenies are widely used in metagenomics analyses, where the goal is exploring what microorganisms exist in an environment by studying their genetic materials. For example, microorganisms can be related to some disease and studying microbiome data can enrich our understanding of such disease [172]. Machine learning techniques are widely used in metagenomics analyses to characterize important microbial communities prevalent in different cohorts, and classifying phenotypes [114, 233, 208].

The accuracy of the ML methods are highly sensitive to the number and distribution of labels in training data. In microbiome analyses, the number of features is commonly much larger than the number of training data points, and hence overfitting of the ML methods is a possibility that can lead to the poor generalization of the model to query samples [45, 46]. When collecting more samples is not feasible, data augmentation techniques have been used to balance the distribution of training data and ameliorate overfitting [46]. For example, existing ML methods such as SMOTE [46] and ADASYN [89] are introduced to address issues related to the unbalanced distribution of labels in the training data, without considering the correlation between features. In Chapter 7, I introduce a new data augmentation algorithm, called *TADA*, that can be used in microbiome analyses to address issues related to limited data size and imbalanced datasets. My algorithm considers phylogenetic relationships between microbial units (ideally, organisms), and a statistical generative model to simulate new samples. Using two biological datasets and traits, and performing different cross-validation experiments I show TADA can improve the performance of ML methods dramatically. This improvement is more pronounced, especially when class labels in the training datasets are highly unbalanced.

In summary, I show phylogenetic reconstruction pipelines are highly sensitive to noise and incongruences omnipresent in genomic data. Moreover, measures for assessing the quality of inferred phylogenies are either not scalable to large datasets or not well-calibrated. In this dissertation, I introduce techniques to reduce noise prevalent in the genomic data that can affect quality of reconstructed phylogenies. Moreover, I present highly scalable methods of inferring, testing, assessing and visualizing phylogenies. Finally, I consider the application of phylogenies in machine learning tasks using metagenomic data. Considering the evolutionary relationship between microbiome features and using a statistical generative procedure, I introduce a new method of augmenting synthetic samples from available training data. This method can be used to address issues of limited sample size in machine learning tasks (e.g., classification or regression). For all methods I explore in this dissertation, I provide amortized running time and show their effectiveness in extensive biological and simulated analyses (if applicable).

Chapter 2

Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction

Inferring species trees from gene trees using the coalescent-based summary methods has been the subject of much attention, yet new scalable and accurate methods are needed. I introduce DISTIQUE, a new statistically consistent summary method for inferring species trees from gene trees under the coalescent model. I generalize my results to arbitrary phylogenetic inference problems; I show that two arbitrarily chosen leaves, called anchors, can be used to estimate relative distances between all other pairs of leaves by inferring relevant quartet trees. This results in a family of distance-based tree inference methods, with running times ranging between quadratic to quartic in the number of leaves. I show in simulated studies that DISTIQUE has comparable accuracy to leading coalescent-based summary methods and reduced running times.

2.1 Background

The evolutionary histories of species and genes can be discordant [147], necessitating a distinction between genes trees and species trees. Incomplete Lineage Sorting (ILS), modeled by the multi-species coalescent (MSC) model [191], is one of the main causes of discordance. A fast approach for estimating the species relationships in the face of such discordances is to first estimate a gene tree for each gene and to summarize the gene trees to build a species tree. The summary method, thus, takes as input a set of gene trees and returns a species tree. A desirable property for a summary method is statistical consistency (a theoretical guarantee that it converges in probability to the correct species tree as the number of error-free genes increases). Many statistically consistent summary methods are available (e.g., ASTRAL [164, 165], BUCKy-population [125], and MP-EST [142]), and coalescent-based species tree estimation is a vibrant field of research, with many recent examples of successful biological analyses [105, 227, 270] (see [75, 161, 160, 180, 229] for criticism of these methods, especially their sensitivity to gene tree error).

Inferring trees using pairwise distances is a well-studied general method of phylogenetic reconstruction [37, 64, 206, 263], and several summary methods are distance-based. These methods first compute a pairwise distance between *species* based on input gene trees and then use a distance method (e.g., neighbor joining [206]) to build the species tree; examples of distance-based summary methods are STAR [143], GLASS [169], NJst [141], and its new implementation, ASTRID [255].

Another powerful general approach to phylogenetic reconstruction is analyzing quartets, which are subsets of four leaves in a tree. Quartet methods first infer a set of quartet trees and then combine them to build a tree on the full dataset [237, 64, 223]. Induced quartet trees have also been used [17, 39, 107, 186, 223] to combine a collection of input trees to build a so-called supertree [27]. Quartet-based phylogeny estimation has been revived in recent years [164, 197, 125, 235, 199] because of its connections to coalescent-based analyses [7, 53, 47]. Under the MSC model, for unrooted species trees with four leaves, the most likely unrooted gene tree is identical to the species tree [7] (but this is not true for larger trees [202, 53]). Furthermore, the length of the internal branch in a quartet species tree (in coalescent units) defines the probabilities of the three possible gene tree quartet topologies [53]. Some recent and statistically consistent quartet-based species tree estimation methods rely on these results. For example, ASTRAL seeks the species tree with the maximum number of quartet trees shared with input gene trees [164, 165].

in this dissertation, I introduce a new coalescent-based summary method, called DISTIQUE (Distance-based Inference of Species Trees from Induced QUartet Elements). Like ASTRAL, DISTIQUE is based on quartets, but instead of directly optimizing a quartet score, it uses quartets to compute pairwise distances, which are then used as input to a distance method. The innovative aspect of DISTIQUE is its method of calculating distances. It chooses two arbitrary "anchor" species and computes the frequency of quartet trees induced by gene trees that include the two anchors as sisters. I show that these frequencies can be transformed into an asymptotically additive distance matrix; using this matrix with a consistent distance-based method (e.g., neighbor joining) gives a statistically consistent summary method. This method would generate a species tree on all species except the two anchors in $\Theta(n^2k)$ (for n species and k gene trees). However, using multiple anchor pairs can increase accuracy and can ensure all species are included in the final tree. Various strategies for choosing anchors and combining their results are introduced, with running times ranging between $\Theta(n^2k)$ and $\Theta(n^4k)$.

After describing DISTIQUE, I show that the anchoring approach can be generalized to any tree inference problem. Assume I have a way to compute the topology and the internal branch length for any quartet of leaves. I show that as long as this quartet estimator is consistent, my anchoring mechanism and a certain family of transformations can be used to compute an additive distance matrix, which in turn can be used to infer the correct tree topology but not correct branch lengths. This result is rather surprising because, for any pair of anchors and a pair of other leaves, the quartet internal branch length will often be very different from the distance between non-anchor leaves. Thus, anchoring produces incorrect pairwise distances that are nevertheless additive for the correct tree topology. DISTIQUE uses anchoring because for the MSC-based species tree inference, pairwise species distances are not straightforward to define but inferring quartet trees is easy. I evaluate the accuracy of DISTIQUE on simulated and biological data and show that its accuracy is competitive with the best alternative methods even when used with relatively small subsets of all possible anchors.

2.2 Methods

Notation and background: Let $\mathcal{L}$ denote the leaf-set of size n. For an unrooted tree T on $\mathcal{L}$, the set of quartet trees induced on all possible $\binom{n}{4}$ quartets of leaves is denoted by Q^T. I use *ab.cd* to denote that a and b are sisters in the quartet tree on $\{a,b,c,d\}$. A tree T is equivalent to a distance matrix D^T, computed by summing lengths of the edges between pairs of leaves, and a distance matrix that corresponds to a tree is called additive [40]. I refer to the unique tree [40] associated with the additive distance matrix D as T^D or T. Also, $T|\mathcal{L}'$ and $D|\mathcal{L}'$ denote T and D restricted to the leaf-set $\mathcal{L}'$.

To test for the additivity of a distance matrix D, I can use the four point condition [40]. For a quartet of leaves $Q = \{a,b,c,d\} \subset \mathcal{L}$, the median and the maximum of the following three values should be the same: $\{D[a,b]+D[c,d], D[a,c]+D[b,d], D[a,d]+D[b,c]\}$. When internal branch lengths are assumed positive, as I do throughout this chapter, the minimum value is strictly smaller than the median. Assuming w.l.o.g. $D[a,b]+D[c,d]$ is the smallest value, I can infer *ab.cd* is the topology induced by T^D. Let $\tau(Q) > 0$ denote the length of the single internal branch in this quartet tree, which I call its “quartet length”; i.e., if $ab.cd \in Q^T$, then $\tau(Q) = \frac{1}{2}(D[a,c] + D[b,d] - D[a,b] - D[c,d])$.

2.2.1 General theoretical results

Definition 1 (Anchored Distances)**.** Given two positive constants α, β and a monotonically increasing function $f(x)$ bounded above by β for positive x (i.e., $0 < f(x) < \beta$ for $x > 0$), two “anchor” leaves $u, v \in \mathcal{L}$, and a tree T equivalent to distance matrix D with the corresponding quartet length function $\tau(Q)$, I define:

$$D'_{uv}[a,b] = \begin{cases} \beta + \alpha.\tau(\{a,b,u,v\}) & ab.uv \notin Q^T \\ \beta - f(\tau(\{a,b,u,v\})) & ab.uv \in Q^T \end{cases} \tag{2.1}$$

$$D'_{v}[a,b] = \sum_{u \in \mathcal{L}-\{a,b,v\}} D'_{uv}[a,b] \tag{2.2}$$

$$D'[a,b] = \sum_{v \in \mathcal{L}-\{a,b\}} \sum_{u \in \mathcal{L}-\{a,b,v\}} D'_{uv}[a,b] \tag{2.3}$$

$$D''[a,b] = \max_{u,v \in \mathcal{L}-\{a,b\}} \max(0, \frac{D'_{uv}[a,b] - \beta}{\alpha}). \tag{2.4}$$

D', D'_u, and D'_{uv} are distance matrices on leaf-sets $\mathcal{L}$, $\mathcal{L}\{v\}$, and $\mathcal{L} - \{u,v\}$, respectively, and are called “all-pairs anchored”, “single anchored”, and “double anchored”. I say D'_{uv} is induced from D anchored by u, v. D'' is called an “all-pairs anchored maximum distance matrix” and is defined on the leaf-set $\mathcal{L}$.

Theorem 1. Let D^T be an additive distance matrix. A double anchored distance matrix D'_{uv} induced from D^T anchored by arbitrary leaves $u, v \in \mathcal{L}$ is an additive distance matrix for the leaf-set $\mathcal{L}' = \mathcal{L} - \{u,v\}$ and corresponds to a tree that is topologically identical to $T|\mathcal{L}'$. Similarly, a single anchored distance matrix D'_v induced from D^T anchored by an arbitrary leaf and an all-pairs anchored distance matrix D' induced from D^T are additive distance matrices for the leaf-sets $\mathcal{L} - \{v\}$ and $\mathcal{L}$, respectively, and correspond to trees that are topologically identical to $T|\mathcal{L}-\{v\}$ and T, respectively.

Theorem 2. An All-pairs anchored maximum distance matrix D'' induced from additive matrix D^T is additive and corresponds to a tree with the identical topology and internal branch lengths to T.

Proof of Theorem 1. Let $\{a,b,c,d\} \subset \mathcal{L}$ be four arbitrary leaves and $\mathcal{L}' = \mathcal{L} - \{a,b,c,d\}$. W.l.o.g assume $ab.cd \in Q^T$. I prove that the four point conditions hold for this arbitrarily chosen quartet in D'_{uv}, D'_v, and D'; I also prove that the four point conditions are true for a tree compatible with the tree T. Proving these conditions for arbitrary quartets completes the proof by results of Buneman [40].

I start with the double-anchored matrix. The four point condition can be written in three ways, but only one of them is compatible with the tree T. Since I assumed w.l.o.g that $ab.cd \in Q^T$, the four point condition compatible with T is:

$$\begin{aligned}\overbrace{D'_{uv}[a,b]+D'_{uv}[c,d]}^{L} &< \overbrace{D'_{uv}[a,d]+D'_{uv}[b,c]}^{R1}\\ &= \overbrace{D'_{uv}[a,c]+D'_{uv}[b,d]}^{R2}.\end{aligned}$$

Proof of Theorem 2: I prove that Equation (2.4) returns the sum of internal branch lengths on the path from a to b on the tree T (I denote this by D^T_{ab}). The theorem immediately follows because a distance matrix compatible with the tree T has to be by definition additive and compatible with it (note that the theorem also claims that D'' gives internal branch lengths). For simplicity, I prove with $\alpha = 1$; extension to other values is simple. If a and b are not sisters in T, there exists an anchor pair (u,v) with quartet topology $au.bv$ and $\tau(a,b,u,v) = D^T_{ab}$; to find such u and v, the following procedure can be followed. Pick u arbitrarily from the sister group of a after rooting T on b and pick v arbitrarily from the sister group of b after rooting T on a. With this choice, it's easy to see that $\tau(a,b,u,v)$ becomes simply the sum

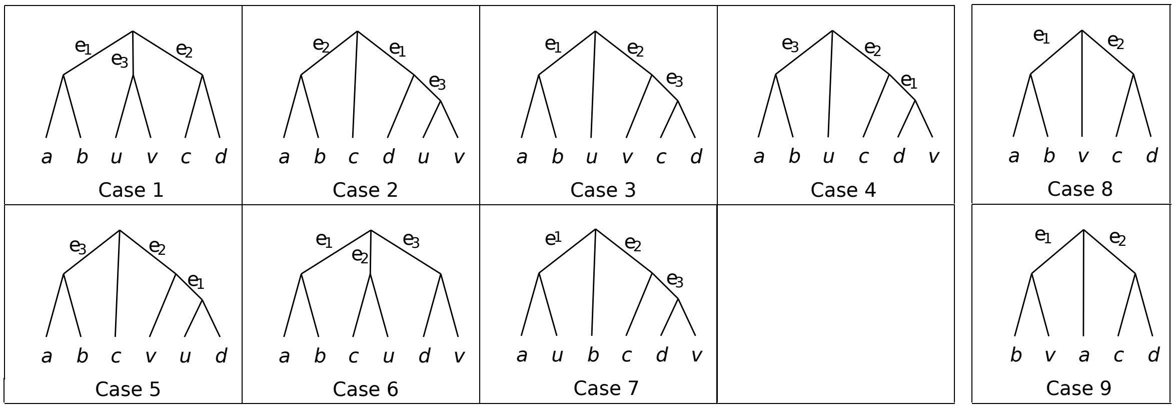


Figure 2.1: Possible ways of adding anchors to a quartet. Left: All 7 possible placements of two anchors u and v on a given quartet topology $ab.cd$. Internal branches are labeled with their length. Right: Placements of a single anchor v on quartet tree $ab.cd$.

of internal branches between a and b; thus, from the first case of Equation (2.1), I have $D'_{uv}[a,b] - \beta = \frac{\tau(a,b,u,v)}{\alpha} + \beta - \beta = D^T_{ab}$. Moreover, $D'_{wz}[a,b] - \beta$ for two other anchors w, z cannot be bigger than D^T_{ab}. That is because if $ab.wz \in Q^T$, then $D'_{wz}[a,b] < \beta$; else, $\tau(a,b,w,z)$ will give the length for a subpath from a to b. Thus, the max function in Equation (2.4) returns D^T_{ab}, as desired. When (a,b) are sisters, $D^T_{ab} = 0$; also $D'_{uv}[a,b] < 0$ for any (u,v), and thus, the max function returns $D''[a,b] = 0$; this is what I want, since for sisters, the length of the internal branch length is zero. Thus, as desired, Equation (2.4) always returns the length of the internal branches in the T between a and b; this completes the proof. □

Figure 2.1 shows all ways of placing anchors $\{u,v\}$ on the quartet tree $ab.cd$. Anchors can be sisters, placed on the internal branch (Case 1) or on a tip branch (Case 2; w.l.o.g, I pick the branch pending to d). When anchors are not sisters, they can be both placed on the internal branch (Case 3), or one on the internal branch and the other on a tip branch (Case 4), or they can be both on terminal branches, which can be done in three ways: u and v can be on the same terminal branch (Case 5), on different but adjacent branches (Case 6), or on non-adjacent branches (Case 7).

In Table 2.1, for each of the seven cases, I compute $L, R1, R2$. I use Equation (2.1) to

Table 2.1: Four point condition for all 7 cases of adding $\{u,v\}$ to a quartet tree, as shown in Figure 2.1 (left side).

	$L = D'_{uv}[a,b] + D'_{uv}[c,d]$	$R1 = D'_{uv}[a,d] + D'_{uv}[b,c]$	$R2 = D'_{uv}[a,c] + D'_{uv}[b,d]$
Case 1	$[\beta - f(e_1 + e_3)]+$ $[\beta - f(e_2 + e_3)]$	$[\beta - f(e_3)]+$ $[\beta - f(e_3)]$	$[\beta - f(e_3)]+$ $[\beta - f(e_3)]$
Case 2	$[\beta - f(e_1 + e_2 + e_3)]+$ $[\beta - f(e_3)]$	$[\beta - f(e_3)]+$ $[\beta - f(e_1 + e_3)]$	$[\beta - f(e_1 + e_3)]+$ $[\beta - f(e_3)]$
Case 3	$[\beta - f(e_1)] + [\beta - f(e_3)]$	$[\beta + \alpha e_2] + [\beta + \alpha e_2]$	$[\beta + \alpha e_2] + [\beta + \alpha e_2]$
Case 4	$[\beta - f(e_3)] + [\beta + \alpha e_1]$	$[\beta + \alpha(e_1 + e_2)]+$ $[\beta + \alpha e_2]$	$[\beta + \alpha e_2]+$ $[\beta + \alpha(e_1 + e_2)]$
Case 5	$[\beta - f(e_2 + e_3)]+$ $[\beta + \alpha e_1]$	$[\beta - f(e_2)] + [\beta + \alpha e_1]$	$[\beta - f(e_2)] + [\beta + \alpha e_1]$
Case 6	$[\beta - f(e_1)]+$ $[\beta + \alpha(e_2 + e_3)]$	$[\beta + \alpha e_3] + [\beta + \alpha e_2]$	$[\beta + \alpha e_2] + [\beta + \alpha e_3]$
Case 7	$[\beta + \alpha e_1] + [\beta + \alpha e_3]$	$[\beta + \alpha(e_1 + e_2 + e_3)]+$ $[\beta + \alpha e_2]$	$[\beta + \alpha(e_1 + e_2)]+$ $[\beta + \alpha(e_2 + e_3)]$

derive $D'_{uv}[x,y]$ values. Where $xy.uv$ is induced by the tree shown in Figure 2.1, I use $[\beta - f(t)]$ and otherwise I use $[\beta + \alpha t]$, where $t = \tau(x,y,u,v)$ is the length of the internal branch for the quartet tree induced by $\{x,y,u,v\}$. For example, for Case 1, $D'_{uv}[a,b] = [\beta - f(e_1 + e_3)]$ because *ab.uv* is induced by the tree, and the length of the edge on the *ab.uv* quartet tree is $e_1 + e_3$; in Case 7, $D'_{uv}[a,b] = \beta + \alpha e_1$ because *ab.uv* is *not* induced by the tree and $\tau(a,b,u,v) = e_1$.

I need to show that $L < R1$ and $R1 = R2$. I remind the reader that all branches are assumed to be strictly positive and that f is a positive and monotonically increasing function bounded from above by β. In all cases, the equality of $R1$ and $R2$ is immediately clear from the Table 2.1. The inequality $L < R1$ follows directly from the fact that $f(x)$ is monotonically increasing in Cases 1, 2, and 5. For Case 3, because of positivity of $f(x)$ and α, I have $L < 2\beta < R$. Similarly, for Case 4, $L < 2\beta + \alpha e_1 < 2\beta + \alpha e_1 + 2\alpha e_2 = R$. Case 6 follows from the positivity of f, and Case 7 is trivially correct for positive branch lengths. Thus, I have shown in all possible relationships between $\{u,v\}$ and the quartet tree, the four point condition holds for the topology consistent with tree T. Therefore, the proof is complete for the double anchored case.

Now consider the "single anchored" distance matrix D^*_v on the leaf-set $\mathcal{L}-\{v\}$ (for a single $v \in \mathcal{L}$). To prove additivity of the single anchored distance matrix, we need to prove the following four point condition:

$$\begin{aligned}
\textstyle\sum_{u\notin\{a,b\}} D'_{uv}[a,b] + \sum_{u\notin\{c,d\}} D'_{uv}[c,d] &< \\
\textstyle\sum_{u\notin\{a,d\}} D'_{uv}[a,d] + \sum_{u\notin\{a,b\}} D'_{uv}[b,c] &= \\
\textstyle\sum_{u\notin\{a,c\}} D'_{uv}[a,c] + \sum_{u\notin\{b,d\}} D'_{uv}[b,d] &
\end{aligned}$$

I divide each sum to terms with $u \in \mathcal{L}'$ and $u \notin \mathcal{L}'$:

$$\begin{aligned}
&\sum_{u\in\mathcal{L}'} D'_{uv}[a,b] + D'_{uv}[c,d] + \\
&\underbrace{D'_{cv}[a,b] + D'_{dv}[a,b] + D'_{av}[c,d] + D'_{bv}[c,d]}_{L} < \\
&\sum_{u\in\mathcal{L}'} D'_{uv}[a,d] + D'_{uv}[b,c] + \\
&\underbrace{D'_{bv}[a,d] + D'_{cv}[a,d] + D'_{av}[b,c] + D'_{dv}[b,c]}_{R1} = \\
&\sum_{u\in\mathcal{L}'} D'_{uv}[a,c] + D'_{uv}[b,d] + \\
&\underbrace{D'_{bv}[a,c] + D'_{dv}[a,c] + D'_{av}[b,d] + D'_{cv}[b,d]}_{R2}
\end{aligned}$$

For $u \in \mathcal{L}'$ terms, the sums are exactly those I analyzed for double anchored distances; thus, the additivity is already proved. Since the sum of two additive distances is additive, it suffices to prove additivity for $u \in \{a,b,c,d\}$ cases, marked as $L, R1$, and $R2$ above.

A single anchor v can be placed (Fig. 2.1) either on the internal branch (Case 8) or on a terminal branch (Case 9) of a quartet tree. I prove $L < R1 = R2$ for these: In Case 8, I have:

$$\begin{aligned} L =&[\beta - f(e_1)] + [\beta - f(e_1)] + [\beta - f(e_2)] + [\beta - f(e_2)] \\ <&4\beta < [\beta + \alpha e_1] + [\beta + \alpha e_1] + [\beta + \alpha e_2] = R1 = R2 \end{aligned}$$

and in case 9,

$$\begin{aligned} L =&2[\beta + \alpha e_1] + [\beta - f(e_2)] + [\beta - f(e_1 + e_2)] < \\ &4\beta + 2\alpha e_1 - f(e_1 + e_2) < \\ &[\beta + \alpha e_2] + [\beta - f(e_1)] + [\beta + \alpha e_1] + [\beta + \alpha(e_1 + e_2)] \\ =&R1 = R2. \end{aligned}$$

Note that the four point condition proved above is for the topology that corresponds to the tree T. The proof for single-anchored distances follows from additivity of sum of additive matrices.

I now prove the additivity for the all-pairs matrix. Equation (2.3) has three types of terms: $\{u, v\} \cap \{a, b, c, d\}$ may have (I) both anchors, (II) one anchor, or (III) none. The four point condition can be written:

$$\begin{aligned} &\overbrace{2D'_{ab}[c,d]}^{I} + \overbrace{\sum_{v \in L'} \sum_{u \in \{c,d\}} D'_{uv}[a,b] + \sum_{u \in \{a,b\}} D'_{uv}[c,d]}^{II} + \\ &\qquad \overbrace{\sum_{u,v \in L'} D'_{uv}[a,b] + D'_{uv}[c,d]}^{III} && < \\ &2D'_{ad}[b,c] + \sum_{v \in L'} \sum_{u \in \{b,c\}} D'_{uv}[a,d] + \sum_{u \in \{a,d\}} D'_{uv}[b,c] + \\ &\qquad \sum_{u,v \in L'} D'_{uv}[a,d] + D'_{uv}[b,c] && = \\ &2D'_{ac}[b,d] + \sum_{v \in L'} \sum_{u \in \{b,d\}} D'_{uv}[a,c] + \sum_{u \in \{a,c\}} D'_{uv}[b,d] + \\ &\qquad \sum_{u,v \in L'} D'_{uv}[a,c] + D'_{uv}[b,d] \end{aligned}$$

For terms of type (III) and (II), the additivity is already proved in double and single anchored cases, respectively. Thus, I need to prove additivity only for terms of type (I), which have no anchors. Let $x = \tau(a,b,c,d)$.

$$\begin{aligned} & 2D'_{ab}[c,d] = 2[\beta - f(x)] < 2\beta < 2[\beta + \alpha x] = \\ & 2D'_{ad}[b,c] = 2D'_{ac}[b,d] \end{aligned}$$

Thus, for all three types, the four point conditions hold for the topology found in T. Proof follows from the fact that the sum of additive terms is additive. □

Theorem 2 is similar to a result given by Brodal *et al.* [36], and is easy to prove. The basic idea is that for any two non-sister leaves $\{a,b\}$, there is a pair of anchors such that in the resulting quartet, a and b are not sisters, and the quartet length is exactly the same as the distance between the two leaves minus their terminal branches. I note that similar to us, Brodal *et al.* use the concept of anchors, but instead of using anchors to *define* distances, they use anchors to efficiently build Buneman trees from *given* distances. Thus, despite some parallels, my anchoring mechanism is novel; In particular, Brodal *et al.* do not prove my surprising result that a *single arbitrarily chosen* pair of anchors gives additive distances for the correct topology.

Theorem 1 states anchored distances induced from an additive matrix will correspond to the same topology as the initial matrix (albeit with wrong branch lengths). This result is surprising, but its usefulness might be less clear. Theorem 1 enables new estimators of the tree topology that rely on quartets to compute pairwise distances. Let $\mathcal{D}$ denote the input data to be used for inferring a phylogeny. Regardless of the nature of $\mathcal{D}$, I require having a quartet estimator. A quartet estimator is a function that given a quartet of leaves Q, uses $\mathcal{D}$ to estimate the quartet tree topology and the quartet length $\tau(Q)$, and is statistically consistent if, as the size of $\mathcal{D}$ increases, the estimated quartet topology and length both converge in probability to correct values. Statistically consistent quartet estimators can be designed for various models (e.g., sequence evolution [234] and the MSC [7, 53]).

Given a statistically consistent quartet estimator, a family of statistically consistent tree inference algorithms can be designed (Algorithm 1). See Section 2.2.2 for details and proofs. The basic idea is the following. I can use the quartet estimator to infer a distance matrix that asymptotically can be made arbitrarily close to an additive distance matrix for the true tree topology. Using a method such as neighbor-joining that infers the correct tree for additive distance matrices with a safety radius will give a consistent estimator of the tree [16].

2.2.2 Tree inference algorithms

Let $\mathcal{D}$ denote the input data (e.g., for a summary method, the set of input gene trees, $\mathcal{G}$, was the input data). Regardless of the nature of $\mathcal{D}$, I require having a quartet estimator:

Definition 2 (Quartet estimator)**.** A quartet estimator $\theta^{\mathcal{D}}(t)$ is a function that given a quartet of leaves $t = \{a,b,c,d\}$, uses $\mathcal{D}$ to estimate the quartet tree topology and the quartet length $\tau(t)$. A quartet estimator is statistically consistent if, as the size of $\mathcal{D}$ increases, the estimated quartet topology and length both converge in probability to correct values.

Given a consistent quartet estimator $\theta^{\mathcal{D}}(.)$, Theorems 1 and 2 define a family of statistically consistent phylogenetic reconstruction methods that range in running time between $\Theta(n^2)$ and $\Theta(n^4)$. Algorithm 1 shows general forms of these algorithms. All-pairs and all-pairs-max, which are the simplest methods in this family, use equations (2.11) and (2.12) to compute the distance matrix. They then compute the tree using neighbor joining [206, 238, 16] (but any consistent distance method with a safety radius [16] could be used).

Theorem 3. All-pairs and all-pairs-max phylogenetic reconstruction methods shown in Algorithm 1 are statistically consistent.

Algorithm 1 Anchored quartet-based algorithms. $\theta^{\mathcal{D}}(t)$ is a quartet estimator and returns the quartet topology and length $\tau(t)$. α and β are constants, and $f(x)$ is a monotonically increasing function bounded above by β for positive x (i.e., $0 < f(x) < \beta$ for $x > 0$). *Anchors*(.) uses some strategy to select a subset of all possible anchor pairs.

```
function ANCHORS($\mathcal{L}$)
    return a set of anchor pairs
function $D'_{uv}(a,b)$
    $(t,d) \leftarrow \theta^{\mathcal{D}}(a,b,u,v)$
    if $t = ab.uv$ then return $\beta - f(d)$
    else return $\beta + \alpha.d$
function $D'(u,v)$
    $D'_{uv} \leftarrow 0_{n-2 \times n-2}$
    for $\{a,b\} \subset \mathcal{L} - \{u,v\}$ do
        $D'_{uv}[a,b] \leftarrow D'_{uv}(a,b)$
    return $D'_{u,v}$
function ALL-PAIRS-MAX($\mathcal{D}$)
    $D'' \leftarrow 0_{n \times n}$
    for $\{a,b\} \subset \mathcal{L}$ do
        for $\{u,v\} \subset \mathcal{L} - \{a,b\}$ do
            if $D''[a,b] < D'_{uv}(a,b)$ then
                $D''[a,b] \leftarrow D'_{uv}(a,b)$
    return NeighborJ($D''$)
function ALL-PAIRS($\mathcal{D}$)
    $D' \leftarrow 0_{n \times n}$
    for $\{u,v\} \subset \mathcal{L}$ do
        $D' \leftarrow D' + D'(u,v)$
    return NeighborJ($D'$)
function DISTANCE-SUM($\mathcal{D}$)
    $\mathcal{M} \leftarrow []$
    for $\{u,v\} \subset$ Anchors($\mathcal{L}$) do
        $\mathcal{M} \leftarrow [\mathcal{M}, D'(u,v)]$
    $DS \leftarrow$ average($\mathcal{M}$)
    return NeighborJ($DS$)
function TREE-SUM($\mathcal{D}$)
    $\mathcal{T} = []$
    for $\{u,v\} \subset$ Anchors($\mathcal{L}$) do
        $\mathcal{T} \leftarrow [\mathcal{T}$,NeighborJ($D'(u,v)$)]
    return SuperTree($\mathcal{T}$)
```

Proof (sketch). Since $\theta^{\mathcal{D}}(.)$ is assumed statically consistent, in limit, it will return the correct quartet topology with arbitrarily high probability, and its estimates of quartet lengths can be made arbitrarily close to true values with any desired probability. From this and Theorems 1 and 2, it follows that distance matrices used in AllPairs and AllPairsMax are arbitrarily close to additive, with high probability. The proof follows from the statistical consistency of neighbor joining for distance matrices that are within a safety radius of additivity [16]. □

Algorithms Distance-sum and Tree-sum use a subset of all $\binom{n}{2}$ anchors, combining the results from multiple anchors using methods I described in Section 2.2.6. Both distance-sum and tree-sum methods first select m pairs of anchors with a criterion of choice (e.g., those mentioned for DISTIQUE in Section 2.2.6 and detailed in Algorithm 5). For each anchor pair $\{u,v\}$, a double anchored distance matrix on $\mathcal{L} - \{u,v\}$ is computed. Tree-sum computes m trees, each on $n-2$ leaves using neighbor joining [206], and then combines the m trees using a supertree method (e.g., SuperFine [244] or MRL [174]). Distance-sum combines distance matrices by averaging them, ignoring missing values. I have not been able to prove consistency for the distance-sum strategy, but tree-sum can be proved consistent with an appropriate choice of the supertree method. Recall a *compatibility* supertree refines all its inputs in the output when the input trees are compatible. I define a set of trees *complete* if every quartet of leaves appears in at least one tree.

Theorem 4. The tree-sum phylogenetic reconstruction method shown in Algorithm 1 is statistically consistent if used with a compatibility supertree and when the anchor selection generates a complete set of m trees.

Proof. By Theorem 1 and arguments similar to those used for Theorem 3, each tree $T_1 \ldots T_m$ on $n-2$ leaves is a statistically consistent estimate; thus, for any $\varepsilon' < 1$, there is a dataset size such that each T_i is correct with probability $1-\varepsilon'$. Taking the largest of these dataset sizes, with probability at least $(1-\varepsilon')^m$, every T_i is correct. By setting $\varepsilon' < 1-(1-\varepsilon)^{\frac{1}{m}}$ I argue that for any ε, there is a dataset size where all T_is are correct with probability at least $1-\varepsilon$. By Lemma 2 the use of a compatibility supertree method and the completeness of m binary and correct trees guarantee that the supertree is the correct tree. □

Combining these algorithms with my approaches for dealing with long branches (i.e., the use of a consensus tree) results in a somewhat more complicated algorithm. Algorithm 2 shows the detailed steps for my default distance-sum approach.

A note on branch lengths: Anchored distances (when computed exactly) set the length of terminal branches to zero. A constant can be added to all distances for inference methods that expect non-zero terminal lengths. AllPairsMax returns statistically consistent estimates of *internal* branch lengths, but branch lengths from my other methods should be ignored.

2.2.3 DISTIQUE (theory)

Problem statement: Given an input dataset $\mathcal{G}$ of a collection of k unrooted gene trees, I seek to find the unrooted species tree topology, assuming gene trees are generated by the MSC model [191].

Next, I first describe anchored distances based on the MSC model used in DISTIQUE. I then describe the algorithmic design of DISTIQUE, including its strategies for selecting anchors, combining results from multiple anchors, and dealing with long branches.

Definition 3 (MSC-based anchored distances)**.** Let $p(ab.uv)$ denote the true probability of observing the quartet topology *ab.uv* in gene trees generated according to the MSC model. I define MSC-based double, single, and all-pairs anchored distance matrices $D^*_{u,v}$, D^*_v, and D^*, respectively on leaf-sets $\mathcal{L}-\{u,v\}$, $\mathcal{L}-\{v\}$ and $\mathcal{L}$ as:

$$D^*_{u,v}[a,b] = -\ln p(ab.uv) \tag{2.5}$$

$$D^*_v[a,b] = \sum_{u\in\mathcal{L}-\{a,b,v\}} -\ln p(ab.uv) \tag{2.6}$$

$$D^*[a,b] = \sum_{v\in\mathcal{L}-\{a,b\}} \sum_{u\in\mathcal{L}-\{a,b,v\}} -\ln p(ab.uv) \tag{2.7}$$

Lemma 1. For species tree estimation under the MSC model, Equation (2.1) simplifies to Equation (2.5) for $\beta = \ln 3$, $\alpha = 1$, and $f(x) = \ln(3-2e^{-x})$. Thus $D'_{uv}[a,b] = D^*_{uv}[a,b] = -\ln p(ab.uv)$.

Theorem 5. Given true quartet probabilities $p(ab.uv)$, D^*_{uv}, D^*_v, and D^* become additive distance matrices that correspond to the true species tree topology on leaf-sets $\mathcal{L}-\{u,v\}$, $\mathcal{L}-\{v\}$, and $\mathcal{L}$, respectively.

From Lemma 1, it follows that Equation (2.5) is a special case of Equation (2.1); Theorem 5 follows directly from Theorem 1.

It may be surprising that D^*_{uv}, which is a special case of D'_{uv}, depends only on quartet topologies and not branch lengths. To see why, readers should recall that p is the quart frequency in *gene trees*, and relates to both the quartet topology and the quartet length in the *species tree*.

True quartet probabilities are not known. Instead, I empirically use $\overline{p}(ab.uv) = \frac{1}{k}|\{t : \mathcal{G}|ab.uv \in Q^t\}|$. Empirical frequencies inferred from gene trees converge in probability to true values as the number of genes increases; thus, it is easy to show (proof omitted):

Corollary 6. D^*_{uv}, D^*_v, and D^* computed using empirical frequencies in a random sample of error-free gene trees converge in probability to an arbitrarily small radius of an additive matrix identical in topology to the true species tree; a consistent distance method with a safety radius [16] run on these matrices is a consistent estimator of the species tree topology.

Computing anchored matrices require $\Theta(n^2k)$, $\Theta(n^3k)$, and $\Theta(n^4k)$ time, respectively for D^*_{uv}, D^*_v, and D^*. Among these matrices, only D^* includes all species.

2.2.4 DISTIQUE (algorithmic design)

DISTIQUE uses double anchored matrices, which can be each computed in $\Theta(n^2k)$. It uses multiple anchors and combines the trees or matrices produced by different anchors. A careful selection of anchors can ensure the final DISTIQUE tree includes all species, and can control its running time between $\Theta(n^2k)$ and $\Theta(n^4k)$. Before presenting my anchoring strategy, I first need to show how DISTIQUE deals with long branches.

Long branches: smoothing and consensus

Smoothing: For species tree branches that are even moderately long, expected frequencies of alternative quartet topologies become exceedingly close to zero. For example, a species tree quartet length of 12 in coalescent units [56] results in a 99.6% chance of observing

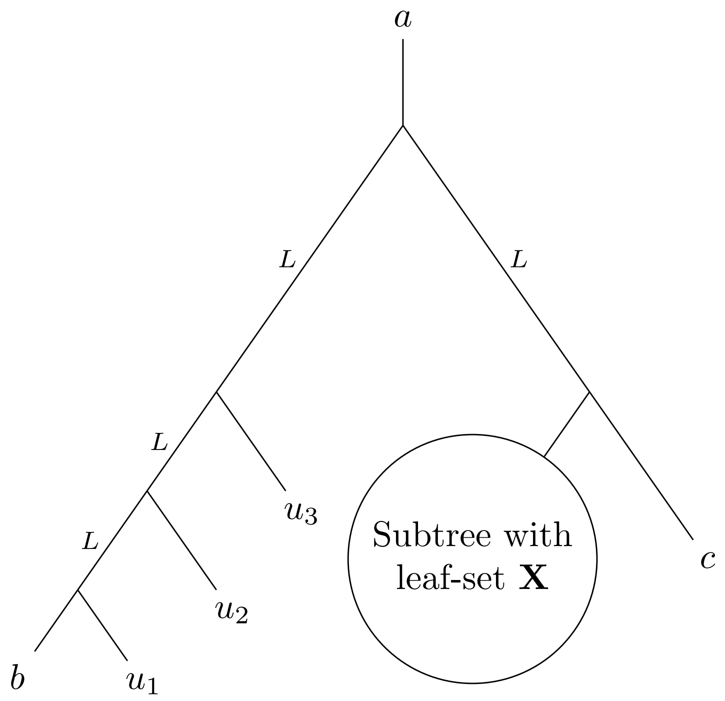


Figure 2.2: An example where long branches can cause problems. See section 2.3 for descriptions.

no discordance among 1000 genes. Thus, my simple empirical frequency estimator $\overline{p}$ can easily be equal to zero, resulting in distances of infinity (Eq. 2.5). To avoid this problem, I use *Krichevsky-Trofimov* [116] (i.e., add-half estimator), which adds a pseudo-count of 0.5 for each of three possible quartet topologies. This estimator has been shown to reach the min-max cumulative loss for KL divergence asymptotically [116].

2.2.5 Computing pseudo-counts for allpairs-max

In allpairs-max methods, in case of zero frequencies for some quartet topologies, without changing the definition of the pseudo-count, the relative information about quartets would be lost. For example, assume I have topology *ab.cz* with long internal branch length, like 16 as mentioned, and *ab.dz* with internal branch length of 20 (longer than the previous length), and no sample for none of the topologies that contradict the species tree. In this case, the distance between a, and c is equal to the distance of a, and d which is $\ln \frac{0.5}{n+1.5}$, where n is the number of samples. So the relative information is lost. In order to avoid this problem, the definition of pseudo count in allpairs-max is slightly changed. In order to have a pseudo count that could capture the relative distances, first, the number of zero quartet topologies are counted. This is called n^0_{ab}. The pseudo count, in this case, is defined as:

$$\ln \prod_{i=1}^{n^0_{ab}} \frac{0.5}{k_i + 1.5} \tag{2.8}$$

Where 0.5 comes from my add half estimator, is probability of zero frequencies, and k_i is the number of samples for quartet topology of i.

2.2.6 Difficulties with long branches and need for majority consensus

In this section, I will explain a problematic case that leads us to use a majority consensus. Figure 2.2 shows an unrooted species tree, with many long branches, with length L. I assume L is long enough that for my given number of gene trees, with high probability, all gene trees will be topologically identical to the species tree. Note that for any number of genes, there are branches long enough where discordance is highly unlikely (I give one example below). Thus, I assume that for branches of length L, I have zero quartet trees that conflict with them.

In my example, the distance of a to c is L and a to b is $3L$ (note that in my analyses, I only calculate branch length for internal branches). I will show that all-pairs can be misled to give a smaller distance for a to b than a to c.

Recall The definition of distance matrices:

$$D'_{uv}[a,b] = \begin{cases} \beta + \alpha.\tau(a,b,u,v) & ab.uv \notin Q^T \\ \beta - f(\tau(a,b,u,v)) & ab.uv \in Q^T \end{cases} \tag{2.9}$$

$$D'_v[a,b] = \sum_{u \in \mathcal{L}-\{a,b,v\}} D'_{uv}[a,b]. \tag{2.10}$$

$$D'[a,b] = \sum_{v \in \mathcal{L}-\{a,b\}} D'_v[a,b] = \sum_{u,v \in \mathcal{L}-\{a,b\}} D'_{uv}[a,b]. \tag{2.11}$$

$$D''[a,b] = \max_{u,v \in \mathcal{L}-\{a,b\}} \max(0, \frac{D'_{uv}[a,b]-\beta}{\alpha}). \tag{2.12}$$

Also recall that for coalescent-based analyses, $\beta = \ln 3$, $\alpha = 1$, $f(x) = \ln(3-2e^{-x})$, and with my use of add-half smoothing, Equation 2.9 simplifies to

$$D'_{uv}[a,b] = -\ln \overline{p}(ab.uv) = -\ln\left(\frac{freq(ab.uv)+0.5}{n+1.5}\right) \tag{2.13}$$

where n is the number of genes, and $freq(ab.uv)$ is the number of genes with induced quartet topology $ab.uv$. Using Equations 2.11 and 2.13 and considering all selections of anchors u and v, I have

$$\begin{aligned} D'[a,b] =& -\binom{3}{1}.\binom{|X|+1}{1}.\ln\left(\frac{0.5}{n+1.5}\right) - \binom{|X|+1}{2}\ln\left(\frac{n+0.5}{n+1.5}\right) \ -\binom{3}{2}\ln\left(\frac{0.5}{n+1.5}\right) \\ &\approx -3(|X|+2)\ln\left(\frac{0.5}{n+1.5}\right) \end{aligned} \tag{2.14}$$

The first term comes from choosing one anchor from $\{u_1, u_2, u_3\}$, and choosing the other anchors from $\{c\} \cup X$. In these cases, a and b are further from each other than the anchors, and because of my assumption about L, the frequency of $ab.uv$ is expected to be zero. The second term comes from choosing both anchors from $\{c\} \cup X$; for these, the frequency of $ab.uv$ is expected to be n. Finally, the last term comes from choosing both anchors from the set $\{u_1, u_2, u_3\}$, where once again, the expected frequency of $ab.uv$ is zero for long enough L, leading to the use of the pseudo count. For large enough n, I can approximate $\ln\left(\frac{n+0.5}{n+1.5}\right) \approx 0$.

The same thing could be written for a and c:

$$\begin{aligned} D'[a,c] = & -\binom{4}{1}.\binom{|X|}{1}\ln\left(\frac{0.5}{n+1.5}\right) - \binom{|X|}{2}\ln\left(\frac{n+0.5}{n+1.5}\right) \quad - \binom{4}{2}\ln\left(\frac{n+0.5}{n+1.5}\right) \\ & \approx -4(|X|)\ln\left(\frac{0.5}{n+1.5}\right) \end{aligned} \tag{2.15}$$

The first term comes from choosing an anchor from of the $\{u_1, u_2, u_3, b\}$, and the other anchor from X; here, for long enough L, frequency of $ab.uv$ would be expected to be zero. The second term comes from choosing both anchors from X, and the last term comes from choosing both anchors from $\{u_1, u_2, u_3, b\}$; in these cases I expect the frequency of $ab.uv$ to be n.

Comparing the Equations 2.14 and 2.15, for $|X| > 2$, the distance between a and b would be smaller than the distance between a and c. This clearly is in contradiction to my tree, so DISTIQUE without the use of majority consensus, in this case, becomes misleading. However, note that all branches with length L are assumed to generate no discordance, and thus will be in the final tree.

Large L: Assume that $L = 16$ (in coalescent units) and I have 1000 gene trees. The probability of having only topologies that agree with the species tree in all 1000 gene trees is $(1-2/3e^{-16})^{1000} = 0.99993$. Thus, I expect all n gene trees to have that topology with very high probability. The probability of having only the species tree topology for branches of length $2L$ and $3L$ is even larger.

Consensus: Smoothing does not fix the larger problem of *distinguishing* between long distances. For example, branches of length 12, 24, or 48 are all very likely to result in no gene tree discordance given 1000 genes; thus, even with smoothing, it remains impossible to distinguish between branches with these very different distances. This limitation makes it impossible to compute distances that reflect the true topology from limited data when the species tree includes adjacent long branches (resembling the saturation problem in phylogenetics [106]). I can construct examples when all gene trees are likely identical, yet my smoothed distances are misleading (Section 2.2.6; Fig. 2.2). However, long branches are easy to recover because they appear in most gene trees. A simple majority rule (50%) consensus of gene trees would return all long branches. Thus, I simply compute the majority consensus and resolve its polytomies using DISTIQUE (Algorithms 2 and 3). Because the majority consensus is proved *not* positively misleading under the MSC [54], my method remains statistically consistent.

To resolve a polytomy, Algorithm 2 first assigns a cluster label to each branch pendant to it, and then builds a tree using DISTIQUE with the cluster labels as leaves; this tree defines a resolution of the polytomy. Given anchor species u,v from two *distinct* clusters, I compute distances between *other* pairs of clusters A and B using Equation (2.5), defining the quartet frequencies as: $\overline{p}(uv.AB) = \frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B}\overline{p}(uv.ab)$. When all clusters in the consensus tree are correct (expected asymptotically), $p(uv.ab)$ values are identical; thus, all $\overline{p}(uv.ab)$ values are empirical estimates of the same true value, and using their average is justified.

Choosing anchors

Algorithm 5 shows DISTIQUE's targeted sampling strategy for choosing a subset of all possible anchor pairs. Let $d_1 \dots d_r$ denote the degree of polytomies in the consensus tree, indexed arbitrarily. For each polytomy i, I randomly partition its d_i clusters into sets of size two; if d_i is odd, I randomly choose a cluster and pair it with the remaining cluster. Then, I randomly choose one species from each cluster. This produces $\lceil\frac{d_i}{2}\rceil$ pairs of anchors for each polytomy i. The total number of anchors is $m = \sum_1^r \lceil\frac{d_i}{2}\rceil = O(n)$ (Lemma 3). Each anchor pair is used to resolve all polytomies on the path between them in the consensus tree. This processes may be repeated several rounds (a user-specified input parameter).

Polytomies of degree 4 or 5 cannot be handled using the double anchored approach because once two clusters are chosen as anchors, only two or three clusters remain which cannot be resolved as unrooted trees. For these small polytomies, I always use all-pairs distance matrices; thus, I choose all $\binom{4}{2}$ or $\binom{5}{2}$ possible pairs of clusters around the polytomy.

Table 2.2: Description of variables and what they contain in Algorithm 2

Variable Name	Variable Type	What Contains
$\mathcal{G}$	list of trees	list of gene trees
$sRnd$	Integer number	# rounds of sampling
$consTree$	tree	consensus tree
$smallPoly$	list	list of small polytomies (degree ≤ 5)
$largePoly$	list	list of large polytomies (degree > 5)
$anchPairsSmallPoly$	dictionary of lists	sampled pairs of anchors around small polytomies
$allAnchPairs$	list	list of all sampled pairs of anchors
$clustPoly$	dictionary of matrices	matrix of taxa under clusters around polytomies.
c_1	index	index of the cluster that the selected anchor belongs to
c_2	index	index of the cluster that the selected anchor belongs to
C	matrix	$C[i][j]$ stores number of quartets of the form $c_1c_2\|ij$
D	matrix	$D[i][j]$ stores maximum possible quartets of the form $c_1c_2\|ij$
$anchDist$	matrix	$anchDist[i][j]$ stores distance between cluster i, and j
$anchNorm$	matrix	$anchNorm[i][j]$ is 1, if $i \notin \{c_1, c_2\}$, and $j \notin \{c_1, c_2\}$. Otherwise is 0.
$totalAnchDist$	matrix	$totalAnchDist[i]$ contains sum of distance matrices for polytomy i
$totalAnchNorm$	matrix	stores # of times distance between j and z for polytomy i estimated
$totalAnchDist$	matrix	averaged distance between i and j for a polytomy
$spTree$	list of trees	$spTree[node]$ estimated species tree for polytomy around $node$
$speciesTree$	tree	Estimated species tree

I need $O(n)$ anchors in this scenario as well (Lemma 3).

Lemma 2. Given m species trees, each on $n-2$ leaves and assuming the set of m trees is complete, and each tree is binary and correct, a compatibility supertree generates the correct tree.

Proof of Lemma 2. Imagine I didn't and the supertree had a wrong internal branch. For every quartet defined around that branch (selecting a leaf from each of the four sides around the branch), by completeness, there is a tree T_i that has that quartet, and because T_i is correct and binary, that quartet should be resolved correctly. Since I can argue this for all quartets around the wrong branch, I get a contradiction where the branch is wrong but all quartets around it are correct. Moreover, since my m trees are binary and complete, the supertree is binary. Imagine it wasn't and take an unresolved quartet around its polytomy; by completeness, that quartet is at least in one T_i. Thus, the supertree does not refine input tree T_i, contradicting the compatibility supertree requirement. Since the supertree has no wrong edges and is binary, it is the correct tree. □

Lemma 3. Given a species tree not necessarily fully resolved with n leaves, anchor-sampling strategy in Algorithm 5 yeilds at most $O(n)$ anchors per each round of anchor sampling.

Proof of Lemma 3. We have two types of polytomies; a polytomy is short if its degree is smaller than 5, and otherwise it is called a large polytomy. My anchoring strategy is that per each round of sampling I have $\lceil \frac{d_i}{2} \rceil$ number of samples for a large polytomy, 6 samples for a polytomy of degree $d_i = 4$, and 10 anchors for a polytomy of degree $d_i = 5$. Note that a species has at most $n-2$ internal nodes, $n-3$ internal branches and n terminal (trivial) branches. Also, each internal branch might be adjacent to at most two polytomies.

$$\sum_{i=1}^{m} d_i \leq 2 \times (n-3) + n = 3n - 6, \tag{2.16}$$

where m is the number of internal nodes. In one extreme, let's consider the case where all the polytomies are large; then, the number of anchors required per each round of anchor sampling is

$$\#anchors = \sum_{i=1}^{m} \lceil \frac{d_i}{2} \rceil \leq \sum_{i=1}^{m} d_i \leq 3n-6, \tag{2.17}$$

which is clearly $O(n)$. Now let's assume that all the polytomies are small. The species tree has at most $n-3$ internal branches which might be in at most two small polytomies. So I might have at most $2(n-3)$ small polytomies, then

$$\#anchors = 2(n-3) \times 10 \tag{2.18}$$

where 10 in above equation is the maximum number of required samples for a short polytomy. Finally, I might have a combination of small polytomies, and large polytomies and the number of required anchors is at most sum of required anchors from Equations 2.17, and 2.18, which is clearly $O(n)$. □

Combining anchors

Once m anchor pairs are selected, DISTIQUE computes m double-anchored matrices and then combines them using one of two methods.

Tree-sum: We first compute m trees, each on $n-2$ leaves using the double anchored method (Corollary 6) and then combine these m trees using a supertree method. Using a *compatibility supertree* (i.e., one that given a set of compatible input trees, outputs a tree that refines all input trees) would make the approach statistically consistent (Theorem 4, SOM). I also use the following approach to filter out outlier anchors. I compute an initial supertree from m anchored trees, then find the RF distance between m trees and the supertree, remove those with an RF distance at least two standard deviations larger than the mean, and recompute the supertree.

Distance-sum: The distance-sum approach creates a summary distance matrix and runs neighbor joining on the summary matrix. The summary distance is simply the average distance of each pair in the set of *m* double anchored matrices. Note that some of the *m* double anchored matrices might not have a value for a given pair of leaves; I treat those as missing values and ignore them when averaging values. The presence of missing values jeopardizes my proofs of statistical consistency.

Let D^*_{uv} and D^*_{wz} be two double anchored matrices produced using two disjoint pairs of anchors. If the two matrices are reduced to the $n-4$ leaves common between them (i.e., $\mathcal{L}' = \mathcal{L} - \{u, v, w, z\}$), I get two matrices that asymptotically converge to an additive matrix for the same tree topology (Corollary 6). The sum of two additive distance matrices that correspond to the same tree topology is also additive for the same topology. Thus, $D^*_{uv}|\mathcal{L}' + D^*_{wz}|\mathcal{L}'$ is asymptotically additive for the correct species tree. This provides a theoretical justification for my distance-sum approach. However, distances between four anchors and other leaves are missing in one of the matrices, and thus, their correct placement cannot be guaranteed.

If all $\binom{n}{2}$ anchors are used, the distance-sum approach becomes equivalent to the all-pairs approach and is provably statistically consistent (Theorem 5). On the other hand, using only two pairs of anchors makes the placement of anchors dependent on averages of two numbers, one of which is missing, a clearly problematic scenario. Choosing an intermediate number of anchors, while insufficient for giving proofs of consistency, clearly reduces the impact of missing values. For example, assume I have *m* anchors and each species is included in at most only one of those anchors. The summary distance between each pair of leaves becomes an average of *m* values, among which at most one may be missing. For large enough *m*, I conjecture that the impact of that single missing value is negligible. In the results section, I provide empirical evidence for this conjecture, but future work should explore theoretical proofs. Due to its superior empirical performance, distance-sum is used by default in the

DISTIQUE (see Algorithm 2 for all details).

Running time analysis:

Using all-pairs or all-pairs-max clearly require $\Theta(n^4k)$ time to build the distance matrix and using the default $O(n^3)$ neighbor joining algorithm [238] would result in $\Theta(n^4k)$ total running time. The running times of tree-sum and distance-sum depend on the selection of anchors, and also the exact distance method and supertree method used. Building each double anchored distance matrix requires $\Theta(n^2k)$; thus, building m matrices requires $\Theta(n^2mk)$. Using a fast neighbor joining algorithm (e.g., FNJ [62], or NINJA [263]), the running time of distance method can be $O(n^2)$. Clearly, any function between $\Theta(n^2k)$ and $\Theta(n^4k)$ can be obtained by adjusting m. DISTIQUE's default strategy requires $O(n)$ anchors and therefore results in $O(n^3k)$ total running time. For the tree-sum approach, the running time of the supertree method needs to be also added. MRL, which I use here, doesn't have running time guarantees, but ML methods tend to have average running time close to $O(n^2)$ [188].

2.2.7 Experimental setup

I use simulated and real datasets to evaluate the accuracy and scalability of DISTIQUE. I measure species tree accuracy using False Negative (FN) rate, which is equivalent to normalized RF distance [134] here because all estimated species trees are fully resolved.

Datasets

For biological analyses, I re-analyzed a dataset of 2022 supergene trees from an avian dataset [161, 105]. I also use three sets of simulated datasets I used before: a 37-taxon mammalian dataset [160], a 45-taxon avian dataset [161], and datasets used for evaluating ASTRAL-II [165]. The first two datasets are based on biological data and have a single species tree topology, whereas the last dataset is simulated using SimPhy [149] and has a

Table 2.3: Empirical statistics of simulated Avian and Mammalian datasets. Model condition $2X$ corresponds to the case where ILS is reduced by increasing the branch lengths (2 times longer), and $0.5X$ represents the case where ILS is increased by reducing the branch lengths (2 times shorter). In the same way, the model condition with $0.2X$ corresponds to the case where ILS is reduced by dividing the branch lengths by five. Average Robinso-Foulds (RF) distances between true gene trees and the model species tree are provided in *AD to species tree*. *# gene trees* shows the number of gene trees that are available for the corresponding dataset and ILS. *#base pairs* represents the number of base pairs, and *# replicates* shows the number of replicates for the corresponding dataset and ILS. In column *Ref.*, the reference paper for each dataset is provided. For the Mammalian with ILS level $0.2X$, *# replicates* 5 and 10 are for the model conditions where *# gene trees* is 3200, and 1600 respectively. Also for the Avian dataset with ILS level $1X$, *# replicates* 10 is only for the model condition with *# gene trees* 2000.

	ILS	AD to species tree	# gene trees	# base pairs	# replicates	Ref.
Mammalian	$2X$	18%	200	500,true	20	[160]
	$1X$	32%	200	500,,true	20	[160]
	$0.5X$	54%	200	500,true	20	[160]
	$0.2X$	79%	100, 200, 400 800,1600,3200	500’true	5, 10, 20	[160]
Avian	$2X$	35%	1000	500,true	20	[161]
	$1X$	47%	200, 500, 1000, 2000	500,true	10, 20	[161]
	$0.5X$	59%	1000	500,true	20	[161]

different species tree per replicate and has heterogeneous parameters. Avian and mammalian datasets enable us to evaluate performance for relatively small numbers of species, varying ILS and the number of genes. The amount of ILS is changed by multiplying or dividing branch lengths by 2 or 5; shorter branches (0.2X and 0.5X) produce more ILS and longer branches reduce ILS (Table 2.3). I create two collections for these datasets, one where I fix the number of genes (200 for mammalian and 1000 for avian) and vary the amount of ILS, and a second collection, where I fix the amount of ILS (to very high or 0.2X for mammalian and default 1X for avian) and vary the number of genes (200 to 3200 for mammalian and 200 to 2000 for avian). The simPhy dataset [165] has two collections, and is simulated to capture the range of reasonable biological datasets. In the simPhy-ILS collection, I fix the number of species to 201 and show three levels of ILS, ranging from moderate (10 million generations) to very high (500K generations), and for each case, I vary the number of genes (50, 200, 1000). For each case, I have 100 replicates, half with a speciation rate of 10^{-6} and the other half with 10^{-7}. In the simPhy-size, I fix ILS to moderate and speciation rate to 10^{-6}, and change the number of species from 10 to 500, with 50 replicates per dataset.

Methods

I compare various versions of DISTIQUE, described below, against each other, and against ASTRAL-II [165], which is a quartet-based method, the ASTRID [255] (a new implementation of the NJst algorithm [141]), which is a distance-based method, and concatenation using RAxML [232] (CA-ML). ASTRAL and NJst are statistically consistent summary methods and, like DISTIQUE, work on unrooted gene trees and species trees (most other approaches such as MP-EST and STAR need rooted input). Also, these two are among the most accurate summary methods [165, 164, 255, 141, 220].

DISTIQUE: I explore variants of DISTIQUE, changing the distance matrix (comparing all-pairs, all-pairs-max, tree-sum, and distance-sum; see Algorithm 7), the number of anchoring rounds (2 to 8), and the use of consensus. To compare to other methods, I use the default distance-sum DISTIQUE (Algorithm 2), with 2 or 8 rounds of anchoring. DISTIQUE is implemented in python and uses the Dendropy library [242] and uses the FastME [128] as its distance method (but I also tested PhyD* [49]).

2.3 Results

2.3.1 Comparison between DISTIQUE variants

I start by comparing all-pairs and all-pairs-max variants, each applied to either the entire set of species or to polytomies of a 50% majority rule consensus (default), limiting my study to the 37-taxon and 45-taxon avian and mammalian datasets where $\Theta(n^4k)$ methods could run. On both datasets, a surprising pattern emerges. Without the use of consensus, the error unexpectedly goes up with decreased ILS, a pattern that is more pronounced for all-pairs-max (Figs. 2.3 and 2.4). As discussed before, I attribute this pattern to difficulties of estimating long quartet lengths. When consensus is used within DISTIQUE, the accuracy improves with decreased ILS, as expected (Figs. 2.3 and 2.4). Depending on the level of ILS, the consensus tree is unresolved for 25% to 95% of branches, leaving much to DISTIQUE to resolve. Overall, all-pairs methods has better accuracy than all-pairs-max, a result that I do not find surprising. Based on these results, hereafter, I only show results for DISTIQUE applied to a majority consensus, and I omit all-pairs-max.

I compared the three algorithms, all-pairs, tree-sum, and distance-sum (the last two with eight rounds of anchor sampling), and observed that the distance-sum is competitive with all-pairs and outperforms tree-sum (Table 2.4). The difference between all-pairs and distance-sum was never more than 1%.

Distance-sum consistently outperformed tree-sum, by as much as 4% in some cases,

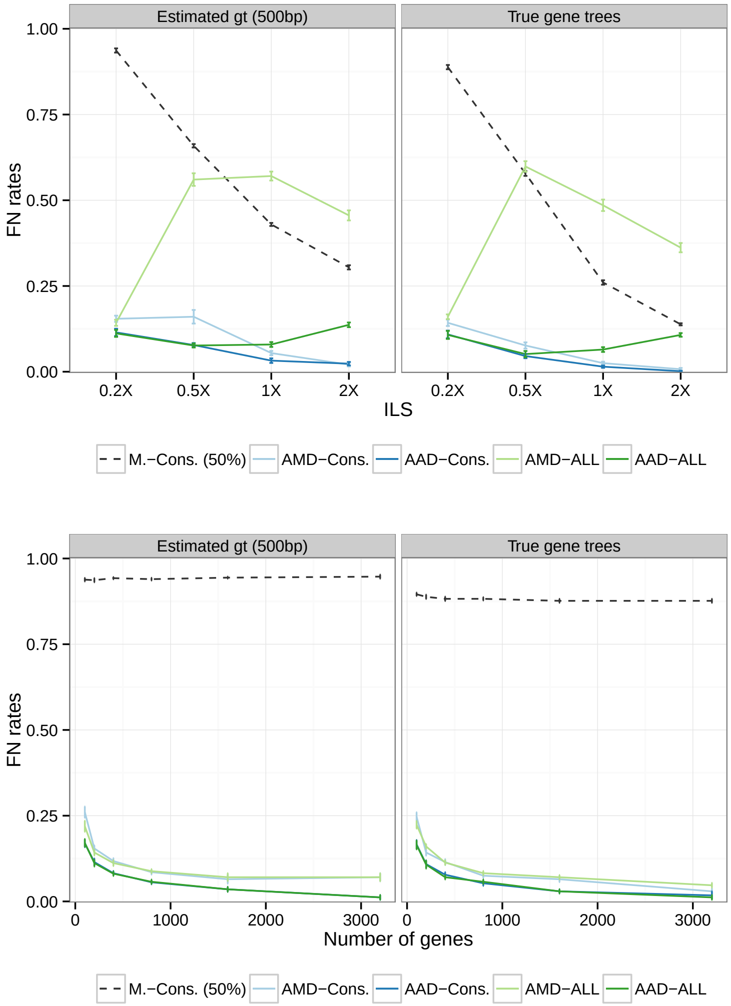


Figure 2.3: Accuracy of different implementations of DISTIQUE for the Mammalian dataset. This figure compares four versions of DISTIQUE: all-pairs-max on the consensus (AMD-Cons.), all-pairs on the consensus (AAD-Cons.), all-pairs-max on the full dataset (AMD-ALL), all-pairs on the full dataset (AAD-ALL). I also show the FN rate of the unresolved consensus tree (black dashed line). (top) number of genes: 200, varying ILS; (below) ILS: 0.2X, varying number fo genes. Mean and standard error of species tree error is shown for true and estimated gene trees (500bp alignments). With very high ILS (0.2X), the accuracy for all of the implementations of DISTIQUE are close. As ILS decreases, when DISTIQUE is applied to the entire dataset, the error goes up, which is more pronounced for all-pairs-max. I attribute this pattern to difficulties of estimating long quartet lengths. When DISTIQUE is used to resolve polytomies in the consensus tree, the accuracy improves with decreased ILS, as expected. Note that even with reduced ILS, the consensus tree on estimated gene trees misses more than 25% of branches, and leaves some polytomies for DISTIQUE to resolve.

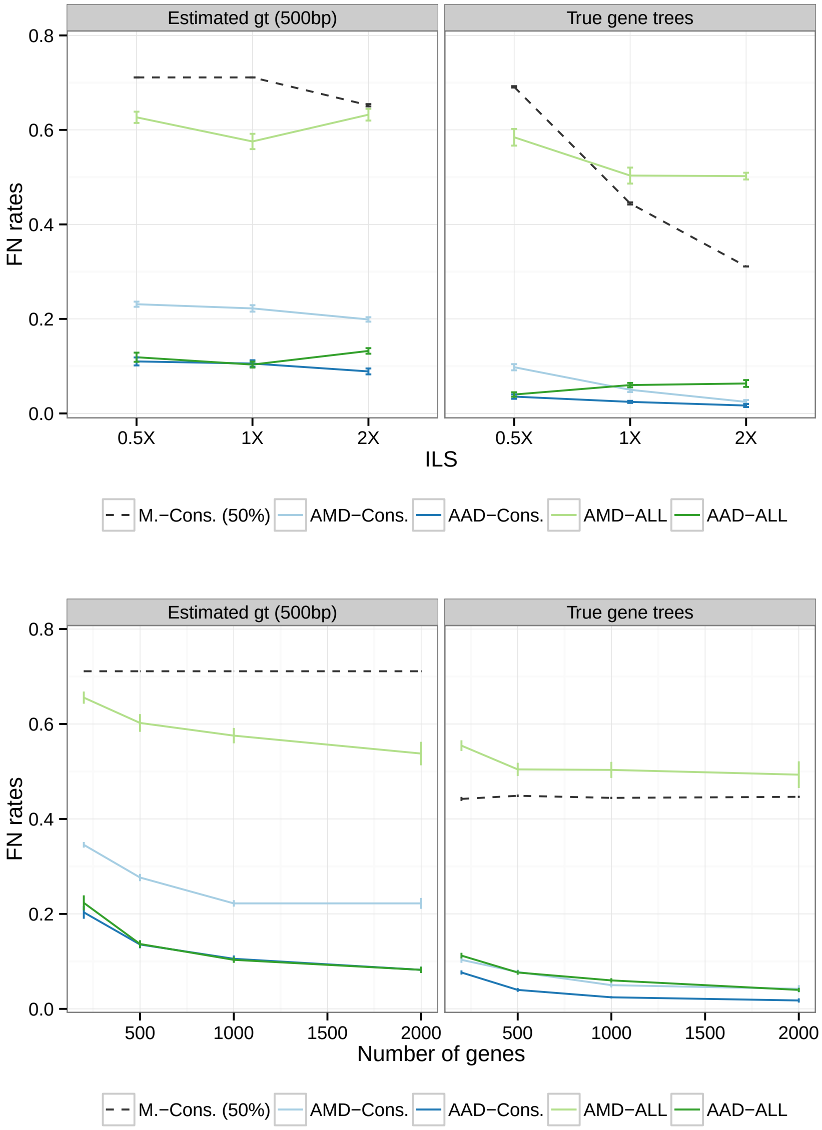


Figure 2.4: Accuracy of different implementations of DISTIQUE for the Avian dataset. This figure compares four versions of DISTIUQE on the Avian dataset. Method labels are as in Fig. 2.3. (top) number of genes: 1000, varying ILS; (below) ILS: 1X, varying number of genes. Mean and standard error of species tree error is shown for true and estimated gene trees (500bp alignments). Note that even with reduced ILS, the consensus tree misses more than 30% of branches, and leaves some polytomies for DISTIQUE to resolve.

Table 2.4: DISTIQUE variants on simulated datasets. Distance-sum and tree-sum are both based on 8 rounds. For simPhy-size, all-pairs could not finish given two days of running time for more than 100 species. Where there is at least 1% difference between methods, the best method is shown in bold.

Dataset	#genes	all-pairs	tree-sum	distance-sum
avian-0.5X	1000	**0.10**	0.11	0.11
avian-1X	1000	**0.08**	0.09	**0.08**
avian-2X	1000	**0.05**	0.08	0.06
mammalian-0.2X	200	**0.11**	0.13	**0.11**
mammalian-0.5X	200	**0.06**	0.12	0.07
mammalian-1X	200	**0.04**	0.08	**0.04**
mammalian-2X	200	**0.02**	0.04	**0.02**
simphySize-10	50	0.03	0.03	0.03
simphySize-10	200	0.02	0.02	0.02
simphySize-10	1000	0.02	0.02	0.02
simphySize-50	50	**0.07**	0.10	**0.07**
simphySize-50	200	**0.04**	0.07	**0.04**
simphySize-50	1000	0.03	0.04	0.04
simphySize-100	50	**0.08**	0.11	**0.08**
simphySize-100	200	**0.05**	0.06	**0.05**
simphySize-100	1000	**0.03**	0.05	0.04

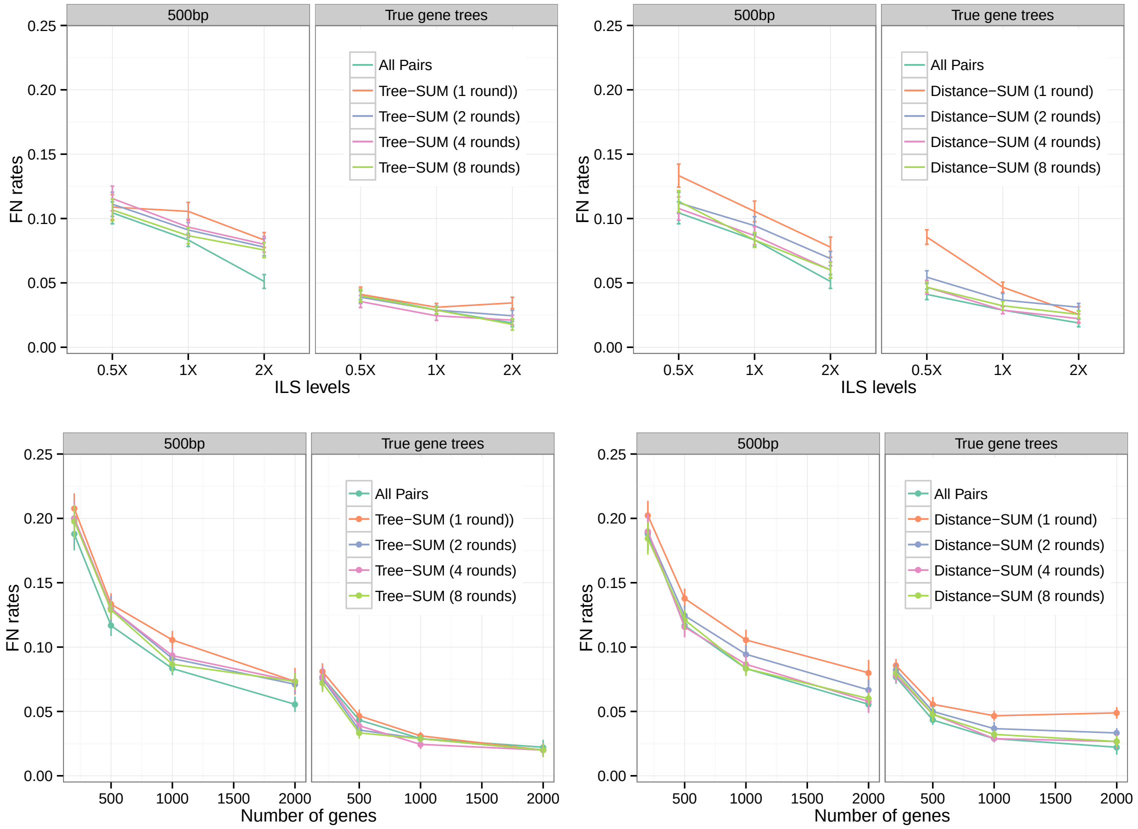


Figure 2.5: Impact of number of rounds of anchor sampling on accuracy. Dataset: Avian simulated dataset with (top) 1000 genes and varying ILS, and (bottom) 1X ILS and varying numbers of genes. The accuracy of DISTIQUE in its $\Theta(n^4)$ allPairs version is compared against $O(n^3)$ double anchored with $m = O(n)$ anchors and both tree-sum and distance-sum strategies of combining multiple anchors with various numbers of rounds of sampling. For tree-sum 2 rounds of sampling is somewhat better than one round, but beyond 2 rounds, little improvements are observed. For distance-sum, further improvements are observed as the number of rounds increases, but four rounds seems to give a reasonably good estimate without increasing the number of sampling rounds too much.

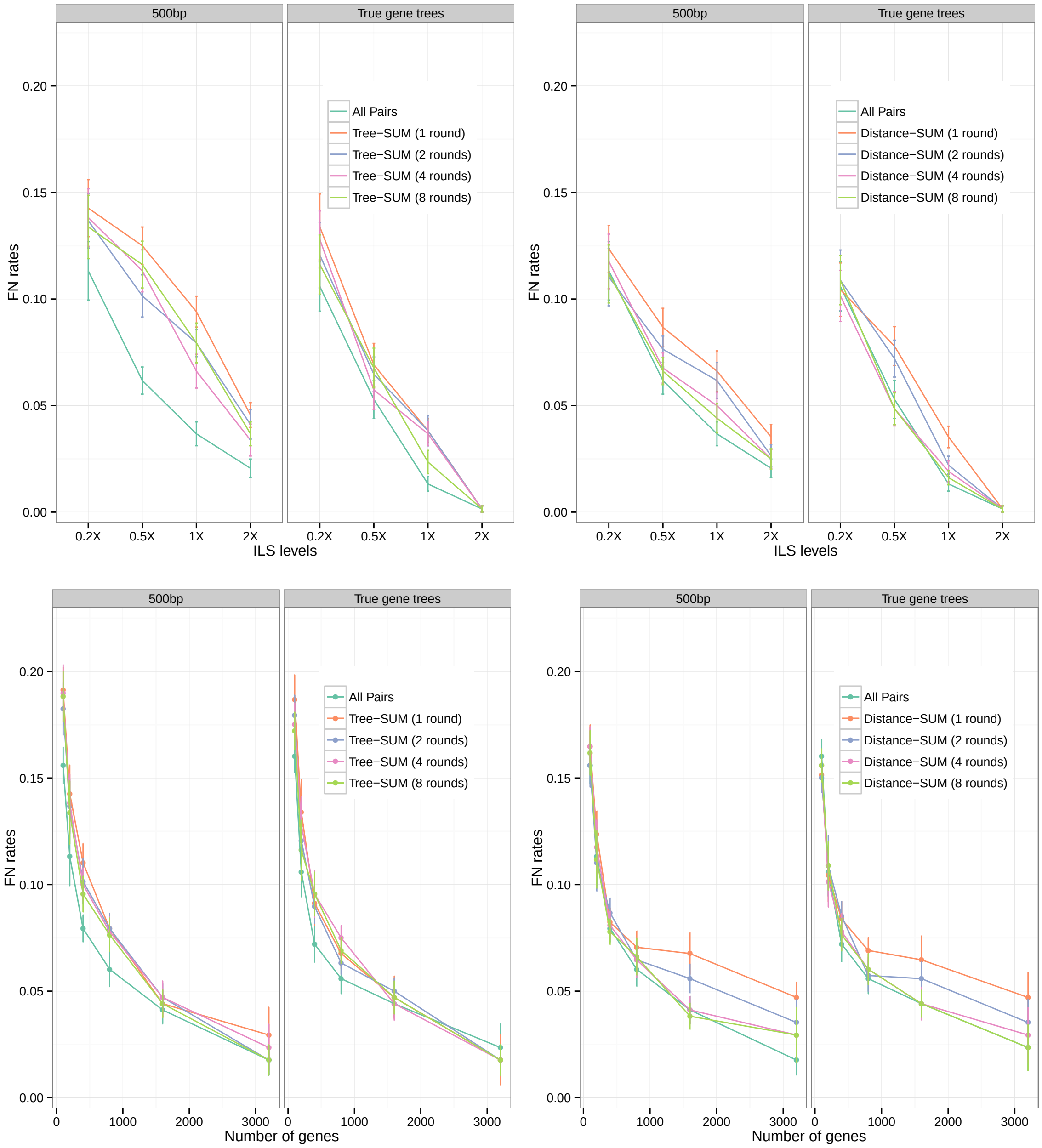


Figure 2.6: Impact of number of rounds of anchor sampling on accuracy. Dataset: Mammalian simulated dataset with (top) 200 genes and varying ILS, and (bottom) 0.2X ILS and varying numbers of genes. The accuracy of DISTIQUE in its $\Theta(n^4)$ allPairs version is compared against $O(n^3)$ double anchored with $m = O(n)$ anchors and both tree-sum and distance-sum strategies of combining multiple anchors with various numbers of rounds of sampling. For tree-sum 2 rounds of sampling is somewhat better than one round, but beyond 2 rounds, little improvements are observed. For distance-sum, further improvements are observed as the number of rounds increases, but four rounds seems to give a reasonably good estimate without increasing the number of sampling rounds too much.

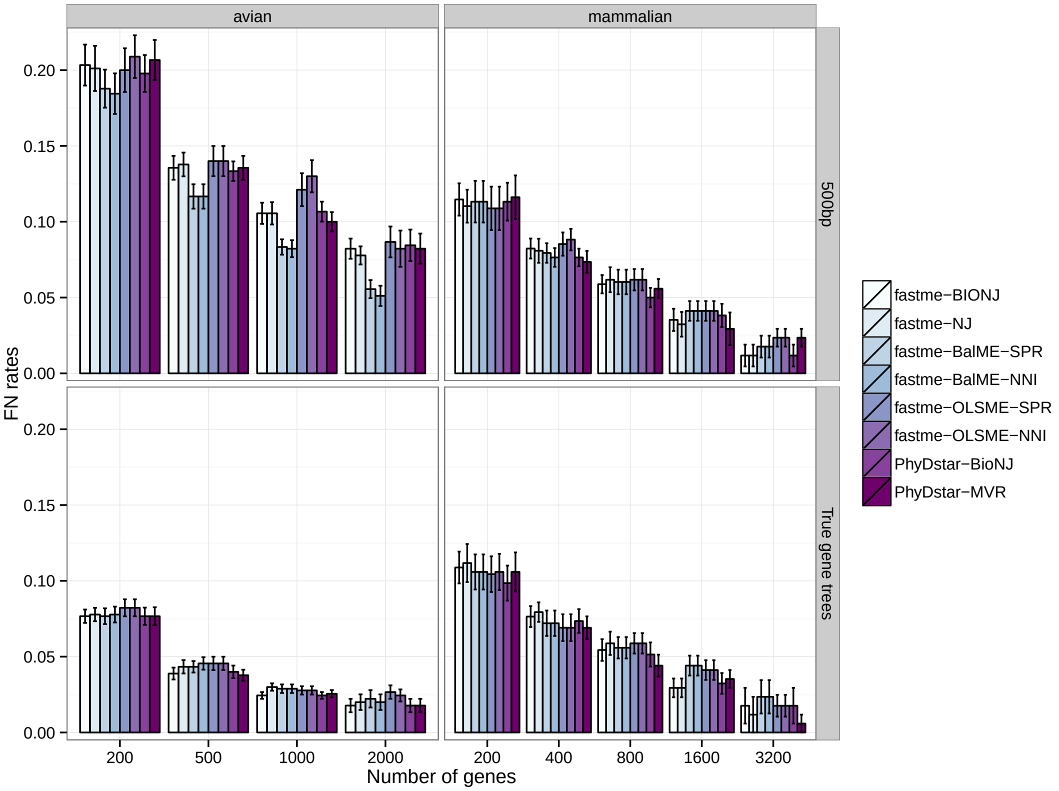


Figure 2.7: Impact of distance method on accuracy. Dataset: Mammalian and aviand simulated dataset with varying levels of ILS on both true and estimated gene trees. Each bar graph is over results from different numbers of genes, ranging between 100 and 3200 for mammalian, and 200 to 2000 for avian. The accuracy of DISTIQUE with various distance methods is used. Methods used from the FastME suit: BioNJ, Neighbor-Joining, BalME (-s), BalME (-nni), OLSME (-s), OLSME (-nni). Methods used from the PhyD* suit: BinoNJ and MVR.

despite the fact that tree-sum is provably consistent and distance-sum has not been proved consistent. Thus, I chose to set the default DISTIQUE implementation to distance-sum.

I next evaluated the impact of anchor sampling by changing the number of rounds of targeted sampling between 1 and 8 on the avian and mammalian datasets (Figs. 2.5 and 2.6). The distance-sum method had substantial improvements when going from one to two rounds, and generally much smaller improvements after that. I show results for both 2 and 8 rounds when comparing DISTIQUE to other methods.

Finally, I checked the impact of the exact distance method used inside DISTIQUE (Fig. 2.7), using a variety of methods implemented inside FastME [128] and PhyD* [49] on both mammalian and avian datasets. There were substantial variations of accuracy among distance methods, especially on the avian dataset. PhyD* tended to have more error, and among methods implemented in FastME, balanced minimum evolution (BME) with SPR moves had the highest accuracy. I chose this option of FastME in the default DISTIQUE.

2.3.2 DISTIQUE versus other methods

SimPhy-size: On this simulated dataset, I compare running time and tree accuracy across methods. Generally, all the methods I studied had similar patterns of accuracy on the simPhy-size dataset, and the mean errors of different methods tended to be within the standard error of each other (Fig 2.8A). According to a two-way ANOVA test with FDR correction [23] for multiple testing ($n = 24$; see caption of Table 2.5) with $\alpha = 0.05$, the error rate of DISTIQUE-8 was statistically indistinguishable from both ASTRAL and ASTRID (Table 2.5). In the few cases where the differences seemed substantial, for example on 500 species and 1000 genes, ASTRAL tended to be the best, followed by both versions of DISTIQUE (but there were exceptions; e.g., 50 species and 1000 genes). Unlike the accuracy, running times of summary methods were quite different (Fig. 2.9). ASTRID was by far the fastest, followed by DISTIQUE-2 and DISTIQUE-8, and ASTRAL was the slowest. For

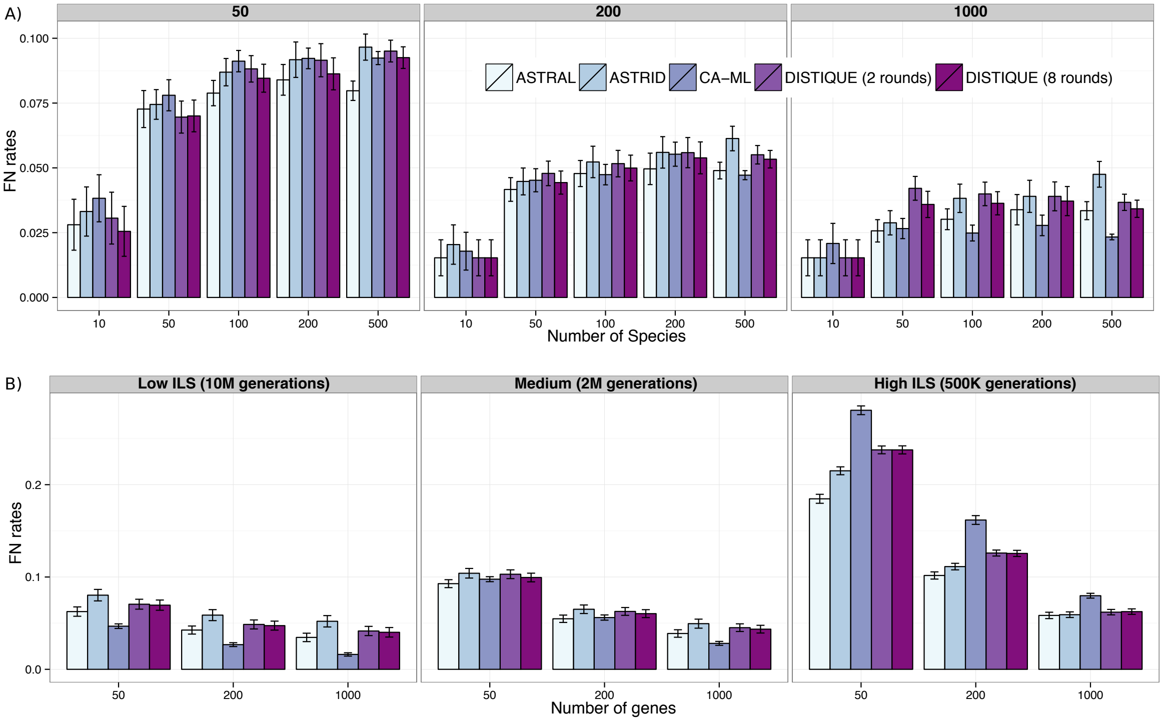


Figure 2.8: DISTIQUE versus other methods on (A) simPhy-size and (B) simPhy-ILS datasets using estimated gene trees. Boxes: (A) number of genes and (B) levels of ILS. The mean and standard error of species tree error are shown over (A) 50 and (B) 100 replicates.

Table 2.5: Two-way ANOVA test with FDR correction. A two-way ANOVA test with FDR correction for multiple testing ($n = 24$) with $\alpha = 0.05$ was conducted to investigate statistically significant differences between different methods. Values bellow $\alpha = 0.05$ (statistically significant values) are shown in bold. Note that column *effects of method* shows the p-values of the impact of the choice of methods (DISTIQUE-8 vs method specified in column *method*). The column *method vs ILS or # species* shows p-values of effects of different levels of ILS (for simPhy-ILS, avian, and mammalian) or different number of species (for simPhy-size) on the relative errors of DISTIQUE-8 and the other method being compared (ASTRID or ASTRAL). The last column shows the effects of varying numbers of gene trees on the relative performance of the methods. FDR correction was applied on all p-values shown in the table (thus, $n = 24$).

dataset	method	effects of method	method vs ILS or # species	method vs # gene trees
avian	astral	0.904	0.191	0.991
avian	astrid	**0.004**	0.904	0.991
mammalian	astral	**0.025**	0.904	0.991
mammalian	astrid	0.904	0.226	0.991
simPhy-ILS	astral	$< \mathbf{1e^{-6}}$	**0.001**	**0.039**
simPhy-ILS	astrid	0.966	**0.001**	0.904
simPhy-size	astral	0.121	0.991	0.991
simPhy-size	astrid	0.137	0.991	0.991

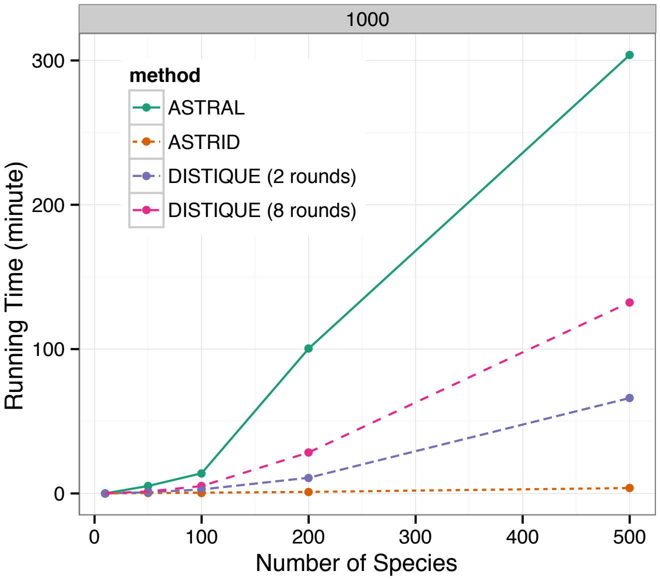


Figure 2.9: Running time comparisons on the simPhy-size datasets with 1000 genes (Fig. 2.10 has other numbers of genes). Lines show the average running times (50 replicates) in hours.

example, with 500 species and 1000 genes, DISTIQUE-2 and DISTIQUE-8 ran in about 1.1 and 2.2 hours, while ASTRAL took 5 hours, and ASTRID took only 7.5 minutes.

SimPhy-ILS: On the simPhy-ILS dataset where the number of species is fixed to 201, differences between various summary methods were generally small (Fig 2.8B), but overall, ASTRAL was significantly better than DISTIQUE-8 ($p < 0.001$). However, DISTIQUE-8 and ASTRID were indistinguishable (Table 2.5). The magnitude of the difference between

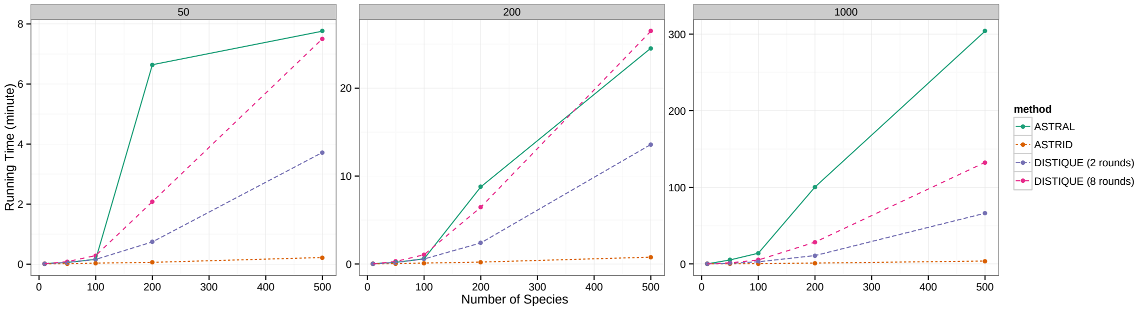


Figure 2.10: Running times of DISTIQUE versus other methods for the simPhy-size dataset. Average running times of ASTRAL, NJst, and DISTIQUE are shown in minutes for different numbers of genes (boxes).

ASTRAL and DISTIQUE-8 significantly depended on the level of ILS ($p = 0.001$), where with low or medium ILS levels, the two methods had a similar error, but with increased ILS, ASTRAL outperformed DISTIQUE; the differences were more pronounced when I had fewer gene trees (significant: $p = 0.039$; Table 2.5).

Avian: On the avian dataset (Fig 2.11A), ASTRID was generally the best method, followed by DISTIQUE-8 (which was significantly worse; $p = 0.004$) and then ASTRAL; CA-ML was the worst. Differences between ASTRAL and DISTIQUE-8 were not statistically significant (Table 2.5). The largest difference between DISTIQUE-8 and the best method was for 0.5X ILS, where DISTIQUE-8 had 2.9% more error than ASTRID.

Mammalian: On this dataset (Fig 2.11B), overall, ASTRAL was the best method, and was significantly better than DISTIQUE ($p = 0.025$). DISTIQUE and ASTRID were overall statistically indistinguishable (Table 2.5). The relative error of concatenation depended on the level of ILS, which was much worse than summary methods for high levels of ILS, but better for low levels of ILS.

2.3.3 Biological results

On the avian dataset, I ran ASTRAL, ASTRID, and DISTIQUE-8 and used both bootstrapping [214] and local posterior probability (pp) [210] to quantify branch support (Figs. 2.12 and 2.13). Bootstrap support was generally high, but the local pp was low for many branches. DISTIQUE and ASTRID differed on three branches. Of these, one, related to the first neoavan split, had high local pp support in ASTRID (0.98) but very low local pp in DISTIQUE; the remaining conflicts had local pp below 0.58 in both trees. ASTRAL and DISTIQUE differed in six branches, and all of these had local pp below 0.58 in DISTIQUE, and all but one also had low local pp (≤ 0.9) in ASTRAL. None of these conflicting relationships have been well resolved in the literature. Interestingly, many of conflicting branches with low local pp had high bootstrap support. It can be argued

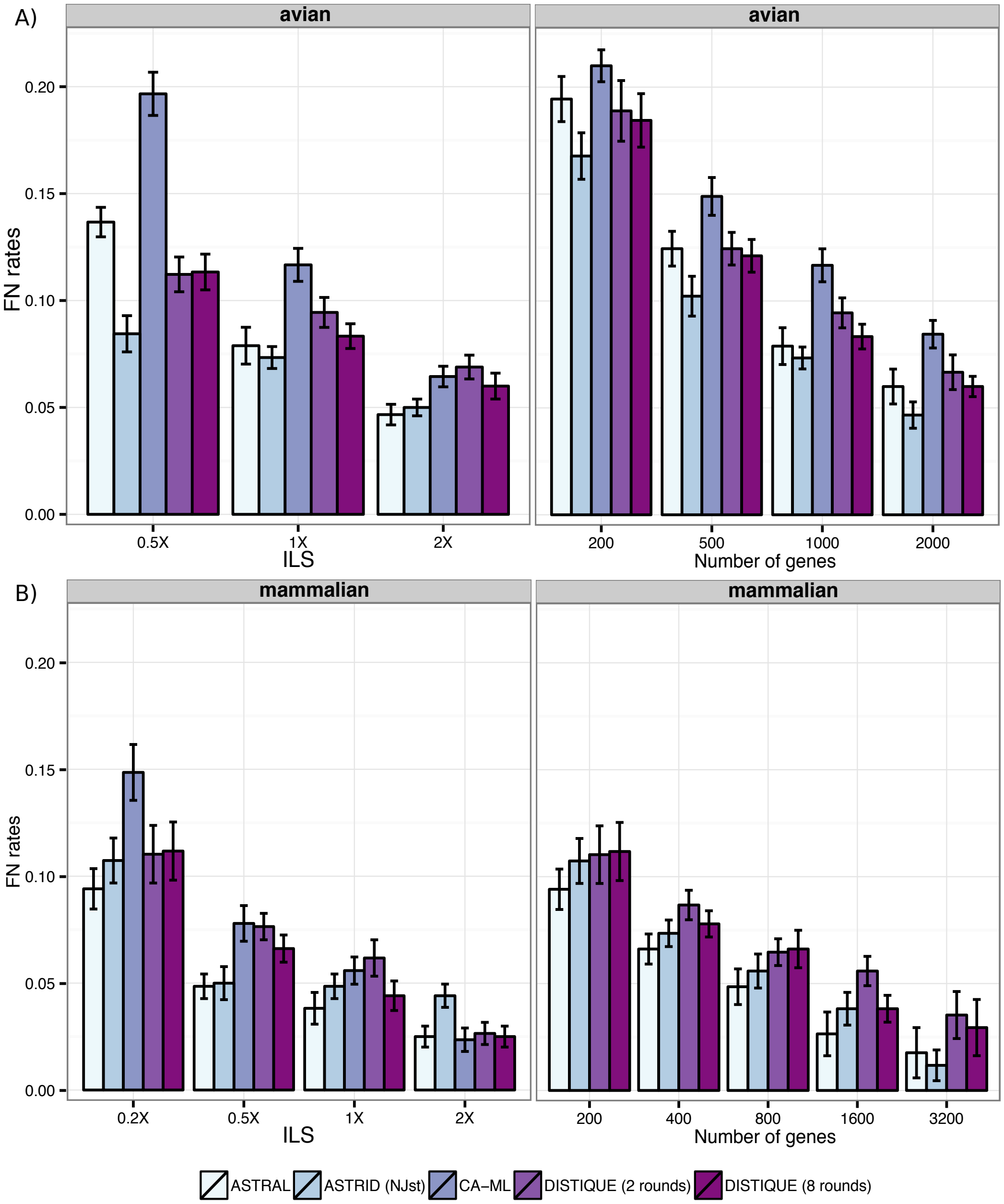


Figure 2.11: The accuracy of methods on Avian (A) and Mammalian (B) datasets using estimated gene trees. Left: number of genes is fixed (1000 for avian, 200 for mammalian) and ILS levels change. Right: ILS level is fixed (default 1X for avian and very high 0.2X for mammalian). I show average and standard error over 20 replicates, except for 1600 and 3200 genes, which have 10 and 5 replicates, respectively. For the mammalian dataset with 0.2X ILS, due to the large number of gene trees, running concatenation was not feasible.

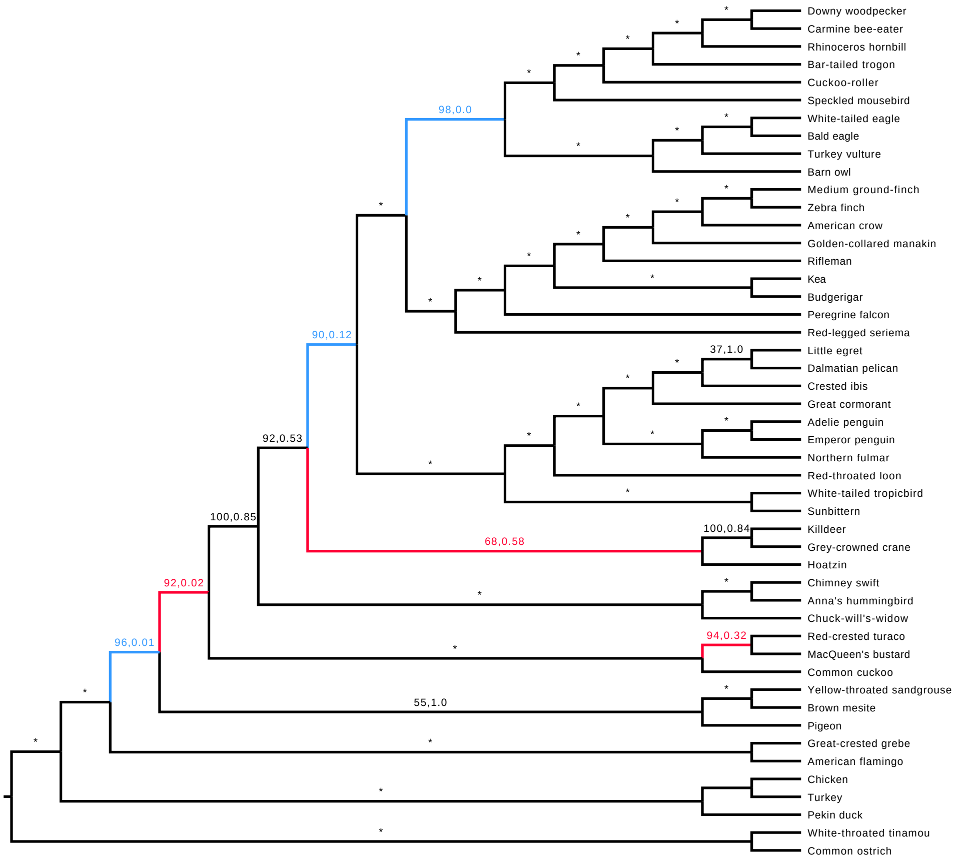


Figure 2.12: Species trees generated using DISTIQUE on Avian biological dataset [214]. Branches that are different between the DISTIQUE tree and both ASTRAL and ASTRID trees are marked in red, while branches that are only different between ASTRAL and DISTIQUE are marked with blue.

that conflicts are due to uncertainties resulting from insufficient data, but bootstrapping misleadingly computes high support [210].

2.4 Discussion

I compared three statistically consistent summary methods, ASTRAL, ASTRID, and DISTIQUE; overall, ASTRAL was at least as good as other methods on most datasets, but ASTRID was occasionally the best. DISTIQUE was often as good as and never more than 3% worse than the best method. The choice of the best method depended on the level of ILS and

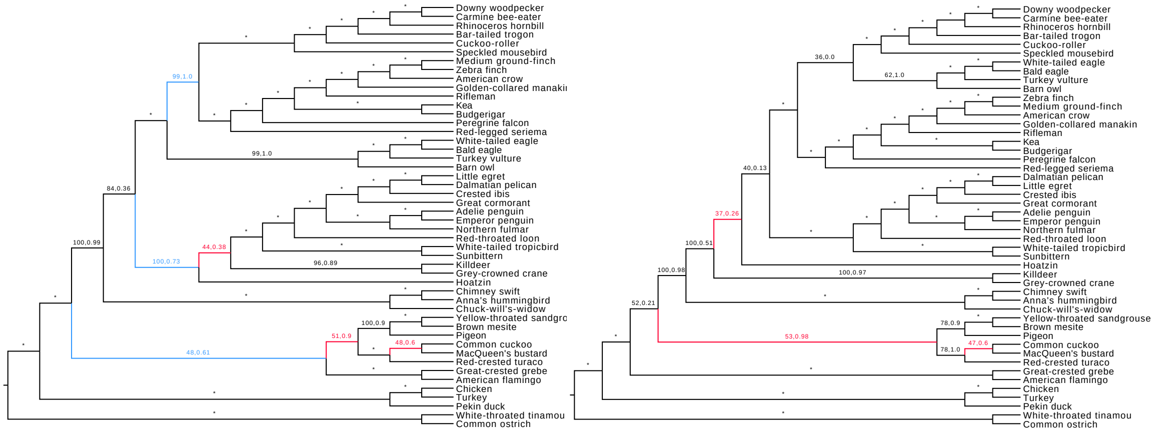


Figure 2.13: Species trees generated using ASTRID (NJst) and ASTRAL on Avian biological dataset [214]. ASTRAL (left) and ASTRID (right) species trees. The branches that are different between DISTIQUE species tree and both ASTRAL and ASTRID species trees are marked with red, while branches that are only different between ASTRAL and DISTIQUE are marked with blue.

the number of genes, suggesting when the level of ILS is expected to be very high, ASTRAL might be the best choice. On the other hand, the running time of DISTIQUE grows more slowly with increased numbers of genes; for datasets with large number of species and tens of thousands of genes, DISTIQUE and ASTRID provide fast alternatives to ASTRAL.

Despite having strong competition in ASTRAL and ASTRID, I believe DISTIQUE is a promising approach, for several reasons. Because of its speed, DISTIQUE can be used for a very fast estimation of species trees, for example, as a starting point for an extensive hill-climbing search. DISTIQUE can also generate a set of trees instead of a single tree, and I plan to study whether these sets of trees can be utilized for defining the search space of ASTRAL.

DISTIQUE is essentially a method for 1) defining distances based on quartets, and 2) subsampling the space of all $\Theta(n^4)$ quartets. The first aspect of DISTIQUE can be replaced by improved ways of defining distances, for example those that better handle gene tree estimation error. Co-estimation of gene trees and the species tree [248] is a computationally challenging problem in general. However, it is reasonable to think that a similar problem defined on quartets, and addressed using distances becomes easier, as some recent theoretical results suggest [51, 199]. DISTIQUE provides a general way for using anchoring introduced in this dissertation to implement novel distance-based gene tree species tree co-estimation in a scalable fashion. Simpler approaches of taking into account gene tree uncertainty, for example weighting various quartets according to coalescent expectations, might also result in improvements. Finally, I note that DISTIQUE's anchoring strategy can be paired with site-based ILS methods such as SVDQuartets [47], and more broadly for other tree inference problems.

2.5 Conclusions

I introduced a general approach for computing tree leaf distances by inferring topologies and internal branch lengths for quartets of leaves. I used my novel anchoring to design DISTIQUE, a new statistically consistent summary method for species tree estimation. DISTIQUE has variants, with several strategies for choosing and combining anchors. The default version of DISTIQUE requires $O(n^3k)$ running time and is much faster than ASTRAL. In terms of accuracy, DISTIQUE was nearly as accurate as ASTRAL with differences that were rarely substantial.

2.6 Acknowledgements

Chapter 2, in full, contains material from SAYYARI, E., AND MIRARAB, S. Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction. BMC Genomics 17, S10 (2016), 101–113. I was the primary investigator and author of this paper.

Algorithm 2 DISTIQUE distance sum algorithm. $\mathcal{G}$ is the set of input gene trees. *sRnd* determines the number of rounds of sampling anchors. The description of variables are provided in Table 2.2. Also the details of functions anchoredFrequencies, distanceAroundPolyNodes, and getClusterAnch are provided in Algorithm 3.

```
function DISTIQUE-DISTANCE-SUM($\mathcal{G}, sRnd$)
    $consTree \leftarrow$ consensus($\mathcal{G}$)
    $consTree \leftarrow$ labelNodes($consTree$)
    $(smallPoly, largePoly, clustPoly) \leftarrow$ findPolytomies($conTree$)
    $polyNodes \leftarrow smallPoly \cup largePoly$
    $(anchPairsSmallPoly, allAnchPairs) \leftarrow$ sampleAnchors($polyNodes, clustPoly, sRnd$)
    for all $node \in largePoly$ do
        for all $anchPair \in allAnchPairs$ do
            $(c_1, c_2) \leftarrow$ getClusterAnch($anchPair, clustPoly[node]$)
            if $c_1 \neq c_2$ then
                $(C, D) \leftarrow$anchoredFrequencies($anchPair, \mathcal{G}, clustPoly[node]$)
                $(anchDist, anchNorm) \leftarrow$distanceAroundPolyNode($anchPair, C, D, clustPoly[node]$)
                $totalAnchDist[node] \leftarrow totalAnchDist[node] + anchDist$
                $totalAnchNorm[node] \leftarrow totalAnchNorm[node] + anchNorm$
    for all $node \in smallPoly$ do
        for all $anchPair \in anchPairsSmallPoly[node]$ do
            $(C, D) \leftarrow$anchoredFrequencies($anchPair, \mathcal{G}, clustPoly[node]$)
            $anchDist \leftarrow$distanceAroundPolyNode($anchPair, C, D, clustPoly[node]$)
            $totalAnchDist[node] \leftarrow totalAnchDist[node] + anchDist$
            $totalAnchNorm[node] \leftarrow totalAnchNorm[node] + anchNorm$
    for all $node \in totalAnchDist$ do
        $totalAnchDist[node] \leftarrow totalAnchDist[node] / totalAnchNorm[node]$
        $spTree[node] \leftarrow$ NeighborJ($totalAnchDist[node]$)
    $speciesTree \leftarrow consTree$
    for all $node \in postorderTraverse(speciesTree)$ do
        if $node \in spTree$ then
            $speciesTree \leftarrow$resolvePoly($speciesTree, spTree[node]$)
    return $speciesTree$
```

Algorithm 3 This algorithm describes details of functions anchoredFrequency, getClusterAnch, and distanceAroundPolyNodes used in Algorithm 2. anchoredFrequency returns frequency of observing quartets of the form $c_1c_2|uv$, where c_1, and c_2 are clusters corresponding to chosen anchors around one polytomy. distanceAroundPolyNode returns distance between cluster u, and v. These distances will be used to infer a resolution of polytomy using phylogeny estimation methods. getClusterAnch determines which clusters anchors belong to. The extension of this algorithm is provided in Algorithm 4.

```
function ANCHOREDFREQUENCIES($anchPair, \mathcal{G}, clusters$)
    $(c_1, c_2) \leftarrow$getClusterAnch($anchPair, clusters$)
    $C \leftarrow []$
    $D \leftarrow []$
    for $\{i, j\} \in$ twoSubsets($\{1, 2, \ldots, |clusters|$) do
        if $i \in \{c_1, c_2\} \vee j \in \{c_1, c_2\}$ then
            $D[i][j] \leftarrow 0$
        else
            $D[i][j] \leftarrow |clusters[i]| \times |clusters[j]| \times |\mathcal{G}| + 1.5$    ▷ 1.5 is for smoothing by add-half
            $C[i][j] \leftarrow 0.5$                              ▷ 0.5 is for smoothing using add-half estimator
    for $g \in \mathcal{G}$ do
        $g \leftarrow$reroot($g, anchPair[0]$)
        $node \leftarrow anchPari[1]$
        repeat
            $pre \leftarrow node$
            $node \leftarrow node.parent$
            $children \leftarrow node.children$
            $children \leftarrow children - pre$
            for $k \in \{1, 2, \ldots, |children|\}$ do
                $C_k \leftarrow$getTaxaUnderClusters($children[k], clusters$)
                for $\{i, j\} \in$ twoSubsets($\{1, 2, \ldots, |clusters|$) do
                    $C[i][j] \leftarrow C[i][j] + C_k[i] \times C_k[j]$
                if $|neighbors| \geq 2$ then
                    for $z \in \{k+1, \ldots, |children|\}$ do
                        $C_z \leftarrow$getTaxaUnderClusters($children[z], clusters$)
                        for $\{i, j\} \in$ twoSubsets($\{1, 2, \ldots, |clusters|$) do
                            $D[i][j] \leftarrow D[i][j] - C_z[i] \times C_k[j] - C_z[j] \times C_k[i]$
        until $node.parent == anchPair[1]$
    return $(C, D)$
```

Algorithm 4 Extension of Algorithm 3

```
function DISTANCEAROUNDPOLYNODE($anchPair, C, D, clusters$)
    for $\{i, j\} \in$ twoSubsets($\{1, 2, \ldots, |clusters|\}$) do
        if $\neg(anchPair[0] \in clusters[i]$ or $anchPair[1] \in clusters[i])$ then
            if $\neg(anchPair[0] \in clusters[j] \vee anchPair[1] \in clusters[j])$ then
                $anchDist[i][j] \leftarrow -\ln(C[i][j]/D[i][j])$
                $anchNorm[i][j] \leftarrow 1$
                $anchDist[j][i] \leftarrow -\ln(C[i][j]/D[i][j])$
                $anchNorm[j][i] \leftarrow 1$
    return $(anchDist, anchNorm)$
function GETCLUSTERANCH($anchPair, clusters$)
    $c \leftarrow \{\}$
    for $i \in \{1, 2, \ldots,$ length($clusters$)$\}$ do
        if $anchPair[0] \in clusters[i]$ then
            $c \leftarrow c \cup clusters[i]$
    return $c$
```

Algorithm 5 Sample anchors algorithm. The variable *polyNodes* contains the polytomies in the tree, and *clustPoly* contains labeled nodes around each polytomy and the list of taxa under each cluster. Also the variable *sRnd* determines the number of sampling rounds around each polytomy.

```
function SAMPLEANCHORS($polyNodes, clustPoly, sRnd$)
    $allAnchors \leftarrow []$
    for all $node \in polyNodes$ do
        $clusters \leftarrow clustPoly[node]$
        if length($clustPoly[node]$) $\geq 6$ then
            $largeAnch \leftarrow$ sampleAnchLargePoly($clusters, sRnd$)
            $allAnchors \leftarrow allAnchors \cup largeAnch$
        else
            $anchSmallPoly[node] \leftarrow$sampleAnchSmallPoly($clusters$)
            $allAnchors \leftarrow allAnchors \cup$ anchSmallPoly[node]
    return $(anchSmallPoly, allAnchors)$
function SAMPLEANCHLARGEPOLY($cluster, sRnd$)
    $i \leftarrow 0$
    $anchPairs \leftarrow []$
    $C \leftarrow \{1, 2, \ldots$ ,length($cluster$)$\}$
    while $i < smpRnd$ do
        $setC \leftarrow C$
        while length($setC$)$> 0$ do
            if length($setC$)$> 1$ then
                $\{u, v\} \leftarrow$ randChooseTwoLeaves($setC$)
            else
                $\{u\} \leftarrow$ randChooseLeave($C - setC$)
                $\{v\} \leftarrow setC$
            $\{taxon1\} \leftarrow$ randChooseLeave($cluster[u]$)
            $\{taxon2\} \leftarrow$ randChooseLeave($cluster[v]$)
            $anchPairs \leftarrow anchPairs \cup \{(taxon1, taxon2)\}$
            $setC \leftarrow setC - \{u, v\}$
    return anchPairs
function SAMPLEANCHSMALLPOLY($cluster$)
    $anchPairs \leftarrow []$
    for $i \in \{0, 1, \ldots$,length($cluster$)$\}$ do
        for $j \in \{i + 1, \ldots$,length($cluster$)$\}$ do
            $\{taxon1\} \leftarrow$randChooseLeave($cluster[i]$)
            $\{taxon2\} \leftarrow$randChooseLeave($cluster[j]$)
            $anchPairs \leftarrow anchPairs \cup \{(taxon1, taxon2)\}$
    return anchPairs
```

Chapter 3

Fast coalescent-based computation of local branch support from quartet frequencies

Species tree reconstruction is complicated by effects of Incomplete Lineage Sorting (ILS), commonly modeled by the multi-species coalescent model. While there has been substantial progress in developing methods that estimate a species tree given a collection of gene trees, less attention has been paid to fast and accurate methods of quantifying support. in this dissertation, I propose a fast algorithm to compute quartet-based support for each branch of a given species tree with regard to a given set of gene trees. I then show how the quartet support can be used in the context of the multi-species coalescent model to compute i) the local posterior probability that the branch is in the species tree and ii) the length of the branch in coalescent units. I evaluate the precision and recall of the local posterior probability on a wide set of simulated and biological datasets, and show that it has very high precision and improved recall compared to multi-locus bootstrapping. The estimated branch lengths are highly accurate when gene tree estimation error is low, but are underestimated when gene tree estimation error increases. Computation of both the branch length and the local posterior probability is implemented as new features in ASTRAL.

3.1 Introduction

The multi-species coalescent model (MSC) of [191] has emerged as the standard method used for reconstructing species trees in the presence of gene tree discordance due to Incomplete Lineage Sorting (ILS) [147, 55]. Many methods have been developed to estimate species trees under the MSC [90, 47, 161, 38]. The most scalable family of MSC-based methods are based on a two-step process where gene trees are first estimated independently for each gene and are then combined to build the species tree using a *summary method*. Many of the summary methods are statistically consistent and thus converge in probability to the true species tree as the number of input error-free gene trees increases; examples of consistent methods include ASTRAL [164, 165], BUCKy-population [125], GLASS [169], MP-EST [142], NJst and ASTRID [141, 255], and STAR [143]. While some methods (e.g., MP-EST) can estimate branch lengths in coalescent units, others only infer the topology. The traditional concatenation approach (where all genes are put together in a supermatrix) can produce high support for incorrect branches [198, 122], and the main goal of statistically consistent summary methods is to address this shortcoming. However, despite the progress in developing methods for species tree reconstruction, little attention has been paid to methods of calculating support.

Bayesian methods [90, 139] readily provide support but remain computationally challenging. Calculating support through bootstrapping [70], while still computationally expensive, is not prohibitively slow and is easily parallelizable. [215] proposed a multi-locus bootstrapping (MLBS) procedure that produces bootstrap replicates by first resampling genes and then sites within those sampled genes. [214] later studied the accuracy of the MLBS approach in the context of distance-based tree reconstruction for 4-taxon trees and explored other strategies where only genes or only sites were resampled. These earlier works did not consider ILS as the cause of discordance; nor did they use summary methods. Nevertheless, the community has adopted MLBS as a standard way of estimating support using ILS-based summary methods; most biological studies using summary methods rely on site-only or site/gene MLBS [227, 264, 105, 189].

Recently, [160] studied the reliability of MLBS support values as a measure of accuracy in simulation studies, and documented both under-estimation and over-estimation of support (for low and high support branches, respectively) using MP-EST [142] and supertree methods such as MRP [190] and MRL [174]. [160] also observed better species tree accuracy when a summary method was run directly on ML gene trees, compared to running the summary method on bootstrapped gene trees first and then taking a consensus. This observation led to the conjecture that MLBS might give biased estimates of the support. Furthermore, [20] found in simulation studies that false positive branches can sometimes have high MLBS support and that many true branches tend to have low support. Obtaining low support for true branches should not be a cause of concern if the lack of support is caused by insufficient data; however, when low support is caused by underestimation, we need better methods of quantifying support.

In this dissertation, I show that using properties of the MSC on four taxa (quartets), I can derive support values that are more precise and more powerful than MLBS, and are much faster to compute. Under the MSC, quartet trees do not have anomaly zones [7, 53], meaning that the most probable gene tree is identical to the species tree for any quartet. Exploiting this property, some summary methods break up gene trees into quartet trees. For example, ASTRAL [165], a summary method used in many recent studies [264, 126, 78, 189, 85, 236, 10], finds the species tree that shares the maximum number of induced quartet trees with gene trees.

I introduce a new method for computing support for species tree branches with regard to a set of unrooted gene trees by calculating Bayesian posterior probabilities. My support values, which I call *local posterior probabilities*, are computed based on gene tree quartet frequencies. For each internal branch of the given species tree, I *assume* that the four *sides* of the branch (Fig. 3.1) are correct, and therefore three topologies are possible around that branch. I introduce a fast algorithm to compute a *quartet support* for each of those three alternatives in $\Theta(nl)$ time (where l is the number of species and n is the number of genes). I then use the quartet support for each alternative topology to derive the posterior probability that it is the correct species topology. Besides producing measures of support, quartet frequencies can be used to derive estimates of internal branch lengths in coalescent units.

My calculation of posterior probabilities is analogous to characterizing a biased die. If I toss a three-faceted biased die n times, my belief in whether the die is biased towards a certain side should depend on the number of tosses and also on the bias of the die (less bias requires more tosses). Similarly, a short branch in the species tree will result in high discordance, and will need many genes to resolve it with high confidence. On the other hand, considering only the MSC and ignoring issues such as long branch attraction, long branches can be easily reconstructed confidently even with few genes.

I show using simulated and empirical datasets that the local posterior probability estimated by my approach is a reliable measure of accuracy. I show that very few highly supported branches are incorrect. Moreover, with a sufficient number of genes, most correct branches have high support. Importantly, I test my methods under conditions where assumptions of my model are violated and show that it remains reliable. My method is available as part of ASTRAL (`https://github.com/smirarab/ASTRAL/`), which now estimates species tree topologies, branch lengths, and local posterior probabilities.

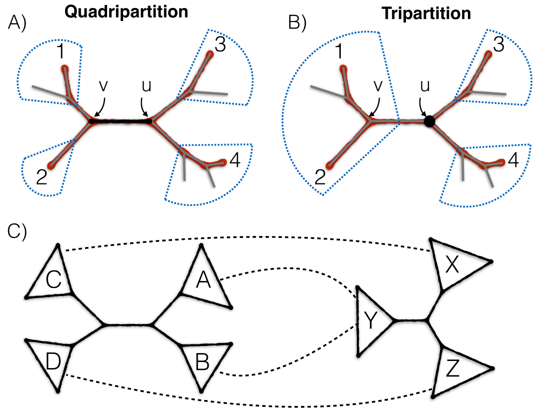


Figure 3.1: Quadripartitions and tripartitions. A) An internal branch (black, in the middle) divides the set of leaves into a quadripartition. A quartet of leaves $\{1,2,3,4\}$ induces a quartet tree (in red) with two internal nodes that map to two nodes in the larger tree (here, u and v). B) An internal node, here u, divides leaves into a tripartition; a selection of two leaves from one side and one from each remaining side gives a quartet mapped to that tripartition. Each quartet tree also maps to a second tripartition (v). C) An example mapping between a quadripartition (e.g., from the species tree) and a tripartition (e.g., from a gene tree); 12 such mappings exist. Note that by finding all quartets of the form $(a,b,c,d); a \in A \cap Y, b \in B \cap Y, c \in C \cap X, d \in D \cap Z$, I can find all quartets around the quadripartition that are mapped to the tripartition with this mapping.

3.2 New Approaches

Definitions: Throughout this chapter, I only consider unrooted trees. Let $\mathcal{L}$ be the set of l leaves (i.e., taxa). Each branch of a tree T creates a bipartition on $\mathcal{L}$, and I say that each of the two partitions is a *cluster* in T. Each *internal* branch divides $\mathcal{L}$ into four clusters, creating a quadripartition (Fig. 3.1). Similarly, an internal node divides $\mathcal{L}$ into three clusters, creating a tripartition (Fig. 3.1). Any quartet q of taxa induces a quartet tree t on T. The two internal nodes of t correspond to two internal nodes in T. When those two internal nodes are the two sides of a single branch in T, I say that the quartet q is *around* that branch. The set of quartets around a given quadripartition can be built by enumerating all selections of one leaf from each of its four clusters.

Problem statement: I am given a set of n gene trees evolved on an unknown true binary species tree according to the MSC model. My aim is to *score* a given internal branch represented as a quadripartition Q to estimate:

1. the probability that Q is in the true species tree, assuming clusters of Q are each correct,
2. the length of Q in coalescent units, assuming Q is a correct branch in the species tree.

Assumptions: I assume evolution is tree-like and true gene trees differ from the species tree only due to ILS, as modeled by the MSC. I also assume I am given an unbiased sample of true gene trees. On real data, I need to instead estimate gene trees from sequence data, and further, it is not always clear that my sample is unbiased, nor that gene trees are generated by the MSC.

Importantly, I further assume that all four clusters around the branch I am scoring are correct. This assumption, which I refer to as the *locality assumption*, makes my computations tractable for large datasets. Similar assumptions have been made in past for fast calculation of local support in the context of maximum likelihood (ML) tree reconstruction from sequence data; e.g., aLRT in PhyML [86] and SH-like support in FastTree-II [188]. Note that to test my method, I use data that violate the locality assumption and I also use estimated gene trees in addition to true gene trees generated by the MSC.

3.2.1 Calculation of local posterior probability

Quartet trees: My approach is based on analyzing quartets defined around the branch Q. For any quartet of leaves around Q, I have three possible topologies, which I will call t_1, t_2, and t_3. In the MSC model, the quartet topology found in the true species tree has the highest probability of appearing in gene trees [7], and the two alternative topologies have identical probabilities. Furthermore, if the total branch length (in coalescent units) between the two internal nodes of the quartet is d, the probability of the dominant quartet topology in gene trees is $\theta = 1 - \frac{2}{3}e^{-d} > \frac{1}{3}$, and the probabilities of both alternative topologies are $\frac{1-\theta}{2} = \frac{1}{3}e^{-d} < \frac{1}{3}$.

I refer to the number of times t_1, t_2, and t_3 are induced in gene trees as *quartet frequencies*, shown as $\bar{n} = (n_1, n_2, n_3)$; note $\sum_1^3 n_j = n$. In the MSC model, conditioned on the species tree, gene trees are independent; Thus, $\bar{n}$ can be modeled as a multinomial random variable $\bar{N}$, with parameters θ, $\frac{1-\theta}{2}$, $\frac{1-\theta}{2}$, where θ corresponds to the species tree topology, and $\frac{1-\theta}{2}$ to the alternative topologies. A similar model is used in the maximum pseudo-likelihood approach of [142] for triplets.

Multiple quartets: There are $m = \prod_1^4 m_i$ quartets around branch Q, where m_i is the size of a cluster of the quadripartition of Q. Note that I can rearrange clusters of Q to obtain two alternative quadripartitions, which I call Q_2 and Q_3. Let $\bar{n}_i = (n_{1i}, n_{2i}, n_{3i}), 1 \leq i \leq m$ be quartet frequencies for all m quartets around branch Q such that n_{1i}, n_{2i}, and n_{3i} correspond to the topologies Q, Q_2 and Q_3, respectively.

Each $\bar{n}_i$ can be modeled as a multinomial random variables $\bar{N}_i$, and $\bar{N}_i$s are identically distributed. To use all $\bar{n}_i$ values, one approach is to assume they are also all independent, and model $(\sum n_{1i}, \sum n_{2i}, \sum n_{3i})$ as observations from a multinomial with $m \times n$ trials. The independence assumption would clearly be incorrect; topologies of different quartets around a branch heavily depend on each other (Fig. 3.2 shows an example). Quartets around a branch are dependent even when the locality assumptions holds. A big problem with the independence assumption is that it inflates confidence because the number of observations (i.e., die tosses) becomes $m \times n$ instead of n, thereby greatly increasing posterior probabilities (note $m \geq l - 3$). Moreover, the dependence of various quartets on each other is intricate and hard to model.

To avoid inflating posterior values by assuming independence, I take the opposite conservative approach. I assume that a hidden random variable $\bar{Z}$ gives a single vector of "true" quartet frequencies around Q and treat each $\bar{N}_i$ as a noisy estimate of $\bar{Z}$. Thus, $\bar{Z}$ follows

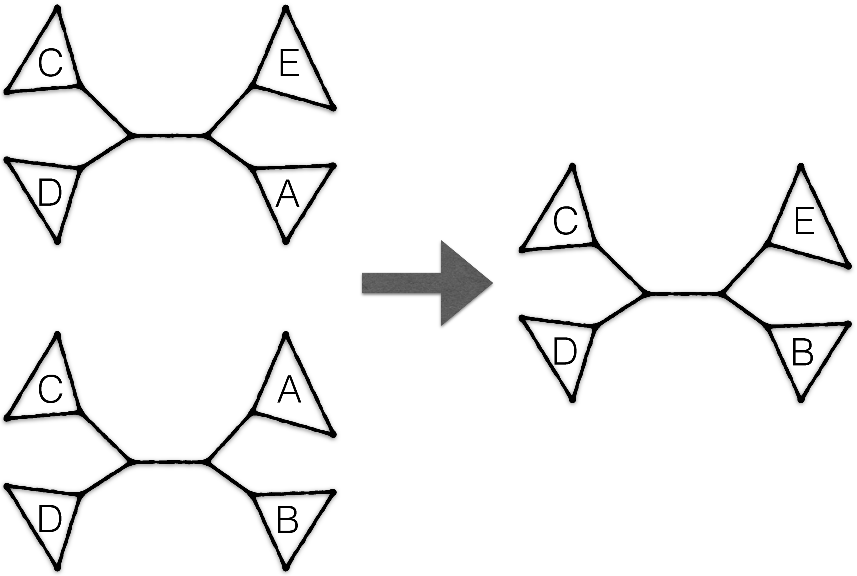


Figure 3.2: Topologies of different quartets around a branch might heavily depend on each other. In this example, species A and B and E are closer to each other than each of them to C or D. Lets assume I observe topology of set of taxa A,B,C,D is $T_1 = A,B|C,D$ in a gene tree G, and the topology of set of taxa A,E,C,D is $T_2 = A,E|C,D$ in G. In this case the topology of set of taxa $T_3 = B,E|C,D$ is completely determined based on topologies T_1, and T_2. So it does not have extra information about the internal branch between them. To see this, imagine putting B on each of the five branches of $A,E|C,D$. Two branches (pending to C and B are not possible, because they contradict the other quartet topology. The other three placements all results in the $B,E|C,D$ topology.

a multinomial distribution with n tries (irrespective of m) and $\bar{N}_i = \bar{Z} + \bar{Y}$ for $1 \leq i \leq m$ where $\bar{Y}$ is a noise term with zero expectation. In the die analogy, I assume the die is tossed n times, and for each toss, I read the outcome m times, each time with some noise. Ideally, I should have a noise model and compute the posterior with respect to the given $\bar{N}_i$ values by marginalizing over $\bar{Z}$. However, a good noise model is not available and the resulting problem becomes hard to solve. Instead, I treat the expected value of $\bar{Z}$ as an observed value, and empirically estimate it by averaging:

$$z_j = \frac{\sum_1^m n_{ji}}{m} \text{ for } j \in \{1,2,3\} \tag{3.1}$$

At the end of this section, I will introduce an efficient $\Theta(nl)$ algorithm to compute $\bar{z}$.

Lemma 4. Let $(\theta_1, \theta_2, \theta_3)$ denote parameters of the true multinomial distribution generating $\bar{Z}$. Note $\sum_1^3 \theta_i = 1$ and the two lower θ_is are identical, and recall z_1 corresponds to the topology of Q.

$$P(\theta_1 > \frac{1}{3} | \bar{Z} = \bar{z}) = \frac{\int_{\frac{1}{3}}^1 P(\bar{Z} = \bar{z} | \theta_1 = t) f_{\theta_1}(t) \mathrm{d}t}{P(\bar{Z} = \bar{z})} \tag{3.2}$$

where f_θ is the prior PDF. The likelihood term is:

$$P(\bar{Z} = \bar{z} | \theta_1 = t) = \Gamma t^{z_1} (\frac{1-t}{2})^{n - z_1} \tag{3.3}$$

where $\Gamma = \frac{\Gamma(n+1)}{\prod_1^3 \Gamma(z_j + 1)}$, and marginal probability is:

$$\begin{aligned} P(\bar{Z} = \bar{z}) &= \sum_{j=1}^3 \int_{\frac{1}{3}}^1 P(\bar{Z} = \bar{z} | \theta_j = t) f_{\theta_j}(t) \mathrm{d}t \\ &= \Gamma \sum_{j=1}^3 \int_{\frac{1}{3}}^1 t^{z_j} (\frac{1-t}{2})^{n - z_j} f_{\theta_j}(t) \mathrm{d}t. \end{aligned} \tag{3.4}$$

Proof. To prove Equation (3.2), one could use Bayes' rule directly,

$$\begin{aligned}P(\theta_1 > \frac{1}{3}|\bar{Z}=\bar{z}) = \frac{P(\theta_1 > \frac{1}{3}, \bar{Z}=\bar{z})}{P(\bar{Z}=\bar{z})} &= \frac{\int_{\frac{1}{3}}^{1} P(\bar{Z}=\bar{z}, \theta_1 = t)\mathrm{d}t}{P(\bar{Z}=\bar{z})} \\ &= \frac{\int_{\frac{1}{3}}^{1} P(\bar{Z}=\bar{z}|\theta_1 = t) f_{\theta_1}(t)\mathrm{d}t}{P(\bar{Z}=\bar{z})}\end{aligned}$$

Equation (3.3) comes from $\bar{Z}$ being a multinomial distributed random variable given $\theta_1 = t$ is fixed and $\theta_1 > \frac{1}{3}$ (so $\theta_2 = \theta_3 = \frac{1-t}{2}$) directly. To prove Equation (3.4),

$$\begin{aligned}P(\bar{Z}=\bar{z}) =& P(\bar{Z}=\bar{z}|\theta_1 > \frac{1}{3}) + P(\bar{Z}=\bar{z}|\theta_1 \leq \frac{1}{3}) \\ =& P(\bar{Z}=\bar{z}|\theta_1 > \frac{1}{3}) + P(\bar{Z}=\bar{z}|\theta_2 > \frac{1}{3} \vee \theta_3 > \frac{1}{3}) \\ =& P(\bar{Z}=\bar{z}|\theta_1 > \frac{1}{3}) + P(\bar{Z}=\bar{z}|\theta_2 > \frac{1}{3}) + P(\bar{Z}=\bar{z}|\theta_3 > \frac{1}{3})\end{aligned}$$

□

Q is in the species tree iff $\theta_1 > \frac{1}{3}$; thus, with Lemma 4 and a prior I can compute the posterior probability.

Prior: In absence of extra reliable information about the species tree topology, which is the most common scenario, the use of an uninformative prior is justified. An uninformative prior would require that the three topologies are equally likely (i.e., $P(\theta_1 > \frac{1}{3}) = P(\theta_2 > \frac{1}{3}) = P(\theta_3 > \frac{1}{3}) = \frac{1}{3}$). Based on Theorem 3.3 from [231], I can prove:

Lemma 5. If the species tree is generated using the Yule process with rate λ, branch lengths are exponentially distributed, and for $t \geq \frac{1}{3}$:

$$f_{\theta_j}(t) = \lambda(3\frac{1-t}{2})^{2\lambda-1} \qquad (3.5)$$

Proof. Under the Yule process with rate λ, the branch lengths D are exponentially distributed with rate 2λ, and the PDF of D is $f_D(x) = 2\lambda e^{-2\lambda x}$ [231]. Note that because of absence of extra reliable information about the species tree topology, I use an uninformative prior, which means all topologies are equally likely to be the dominant one ($Pr(\theta_1 > \frac{1}{3}) = Pr(\theta_2 > \frac{1}{3}) = Pr(\theta_3 > \frac{1}{3}) = \frac{1}{3}$). Knowing that $\theta_j = 1 - \frac{2}{3}e^{-x}$ is in $[\frac{1}{3}, 1)$ as x goes from 0 to ∞ ($j = 1,2,3$), and by using the transformation rule of random variables:

$$f_{\theta_j}(t) = \frac{1}{3}\frac{1}{|\frac{d\theta_j}{dx}|} f_D(x)\Big|_{x=-\ln((1-t)\frac{3}{2})} = \lambda e^{(-2\lambda+1)x}\Big|_{x=-\ln((1-t)\frac{3}{2})} = \lambda((1-t)\frac{3}{2})^{2\lambda-1}. \quad (3.6)$$

□

I use (3.5) throughout the chapter as the prior ($\lambda = \frac{1}{2}$ gives a flat prior). Note that I need that branch lengths in *coalescent units* follow properties of the Yule process; this can be achieved if lengths measured by the number of generations follow the Yule process and N_e is constant for all branches.

Local posterior probability: I now conclude:

Theorem 7. Given 1) a set of n gene trees generated by the MSC on a model species tree generated by the Yule process with rate λ and 2) an internal branch represented by a quadripartition Q where the four clusters around Q are each present in the species tree, let $\bar{z} = (z_1, z_2, z_3)$ be the average quartet frequencies around Q (where z_1 corresponds to the topology of Q); the local posterior probability that the species tree has the topology given by Q is:

$$P(Q|\bar{Z} = \bar{z}) = \frac{h(z_1)}{h(z_1) + 2^{z_2 - z_1}h(z_2) + 2^{z_3 - z_1}h(z_3)} \quad (3.7)$$

for $h(x) = \mathbf{B}(x+1, n-x+2\lambda)(1 - I_{\frac{1}{3}}(x+1, n-x+2\lambda))$. Here, $\mathbf{B}(\alpha, \beta)$ is the beta function, and I_x is the regularized incomplete beta function.

Proof. With locality assumption, $\bar{Z}$ follows a multinomial distribution with parameters $(\theta_1, \theta_2, \theta_3)$. Lack of anomaly zones for unrooted quartets, shown by [6] means that Q is in the species tree iff $\theta_1 > \frac{1}{3}$. Thus, by Lemma 1 I can use (3.2), (3.3), and (3.4) to compute the local posterior probability of Q. By the assumption that (coalescent unit) branch lengths in the species tree are generated by the Yule process, calculation of (3.7) follows from manipulating (3.2), (3.3), and (3.4), as detailed in the following Using Lemma 1,

$$P(\theta_1 > \frac{1}{3}|\bar{Z} = \bar{z}) = \frac{\int_{\frac{1}{3}}^{1} P(\bar{Z} = \bar{z}|\theta_1 = t) f_{\theta_1}(t) \mathrm{d}t}{P(\bar{Z} = \bar{z})}$$

with likelihood term:

$$P(\bar{Z} = \bar{z}|\theta_1 = t) = \Gamma t^{z_1} (\frac{1-t}{2})^{n-z_1}$$

where $\Gamma = \frac{\Gamma(n+1)}{\prod_1^3 \Gamma(z_j+1)}$, and marginal probability:

In these equations $\bar{Z}$ is a multinomial random variable with parameters $(\theta_1, \theta_2, \theta_3)$, $\sum_1^3 \theta_i = 1$ and the two lower θ_is are identical, and recall z_1 corresponds to the topology of Q. Using (3.4), and (3.2),

$$P(Q|\bar{Z} = \bar{z}) = \frac{\int_{\frac{1}{3}}^{1} P(\bar{Z} = \bar{z}|\theta_1 = t) f_{\theta_1}(t) \mathrm{d}t}{\Gamma \sum_{j=1}^{3} \int_{\frac{1}{3}}^{1} g(z_j; n, t) f_{\theta_j}(t) \mathrm{d}t} \tag{3.8}$$

Based on Lemma 2, and Equation 3.3 $f_{\theta_j}(t) = \lambda(\frac{3}{2}(1-t))^{2\lambda-1}$, and $P(\bar{Z} = \bar{z}|\theta_j = t) =$

$\Gamma t^{z_j}(\frac{1-t}{2})^{n-z_j}$. So Equation 3.8 simplifies to:

$$\begin{aligned}P(Q|\bar{Z}=\bar{z}) =&\frac{\int_{\frac{1}{3}}^{1}(\Gamma t^{z_1}(\frac{1-t}{2})^{n-z_1})\lambda(\frac{3}{2}(1-t))^{2\lambda-1}\mathrm{d}t}{\Gamma\sum_{j=1}^{3}\int_{\frac{1}{3}}^{1}(t^{z_j}(\frac{1-t}{2})^{n-z_j})\lambda(\frac{3}{2}(1-t))^{2\lambda-1}\mathrm{d}t}\\ =&\frac{2^{z_1-n}\int_{\frac{1}{3}}^{1}t^{z_1}(1-t)^{n-z_1+2\lambda-1}\mathrm{d}t}{\sum_{j=1}^{3}2^{z_j-n}\int_{\frac{1}{3}}^{1}t^{z_j}(1-t)^{n-z_j+2\lambda-1}\mathrm{d}t}\\ =&\frac{2^{z_1-n}(\int_{0}^{1}t^{z_1}(1-t)^{n-z_1+2\lambda-1}\mathrm{d}t-\int_{0}^{\frac{1}{3}}t^{z_1}(1-t)^{n-z_1+2\lambda-1}\mathrm{d}t)}{\sum_{j=1}^{3}2^{z_j-n}(\int_{0}^{1}t^{z_j}(1-t)^{n-z_j+2\lambda-1}\mathrm{d}t-\int_{0}^{\frac{1}{3}}t^{z_j}(1-t)^{n-z_j+2\lambda-1}\mathrm{d}t)}\end{aligned} \tag{3.9}$$

With $\mathbf{B}(\alpha,\beta)$ as beta function, and I_x as the regularized incomplete beta function,

$$\mathbf{B}(x+1,n-x+2\lambda)=\int_{0}^{1}t^{z_j}(1-t)^{n-z_j+2\lambda-1}\mathrm{d}t$$

and,

$$\mathbf{B}(x+1,n-x+2\lambda)I_{\frac{1}{3}}(x+1,n-x+2\lambda)=\int_{0}^{\frac{1}{3}}t^{z_j}(1-t)^{n-z_j+2\lambda-1}\mathrm{d}t$$

then,

$$P(Q|\bar{Z}=\bar{z})=\frac{h(z_1)}{h(z_1)+2^{z_2-z_1}h(z_2)+2^{z_3-z_1}h(z_3)} \tag{3.10}$$

where

$$h(x)=\mathbf{B}(x+1,n-x+2\lambda)(1-I_{\frac{1}{3}}(x+1,n-x+2\lambda)).$$

□

Examples: The local posterior probability (pp) is a function of the number of genes and the quartet frequency of a branch. As Figure 3.3 shows, a branch that appears in 40% of gene trees has a low 66.1% pp if 50 gene trees given (and alternative topologies are equally frequent); however, with 200 or 500 genes, the same branch will have 93.0% or 99.7% pp, respectively. Thus, a high discordance branch where 60% of genes do not agree with the species tree can still be resolved with high confidence given enough genes. Moreover, the

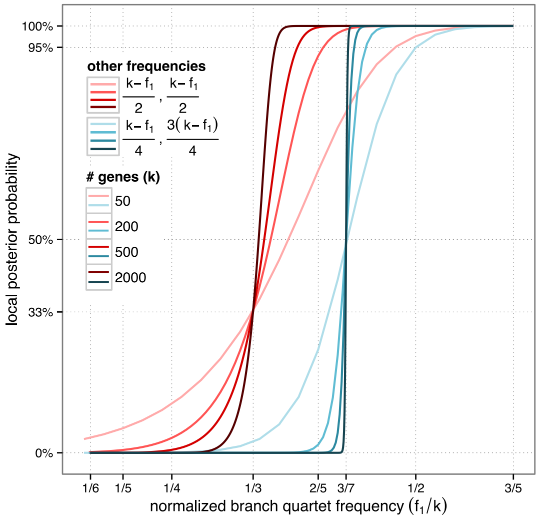


Figure 3.3: The local posterior probability of a branch as a function of its normalized quartet support for varying numbers of genes. Red lines: alternative topologies have equal frequencies (thus, conform to properties of the MSC for x greater than 1/3); Blue lines: alternative topologies don't have equal frequencies (contrary to the MSC).

posterior probability is affected not just by the frequency of the topology being scored, but also by the frequency of the two alternatives. For example, if my branch of interest appears in 40% of gene trees, but a second alternative appears in three-quarters of the remaining genes (i.e., 45% of genes), the branch with 40% frequency has only a 1.90% pp (Fig. 3.3).

3.2.2 Calculation of branch length

Given the true parameter θ for a correct branch, its length in coalescent units [55] is simply $-\ln\frac{3}{2}(1-\theta)$. Thus, I can prove:

Theorem 8. Under conditions of Theorem 7, and assuming the branch represented by Q is in the species tree, the ML estimate for its length is $-\ln\frac{3}{2}(1-\frac{z_1}{n})$ and the MAP estimate is $-\ln\frac{3}{2}(1-\frac{z_1}{n+2\lambda})$ when $z_1 \geq \frac{n}{3}$ and $z_1 \geq \frac{n+2\lambda}{3}$, respectively; otherwise, ML=MAP= 0.

Proof. Assume D, branch length, is a random variable whose range is in $[0\ \infty)$. The ML estimate for the branch length is coming from:

$$d^*_{ML} = \operatorname*{argmax}_{x\geq 0} P_{Z_1|D}(z_1|x;n) \tag{3.11}$$

where $P_{Z_1|D}(z_1|x;n)$ is the likelihood of D. Assuming that Q is the true topology, the likelihood of D is proportional to:

$$P_{Z_1|D}(z_1|x;n) \propto (1-\frac{2}{3}e^{-x})^{z_1}(\frac{1}{3}e^{-x})^{n-z_1} \tag{3.12}$$

computing the log likelihood and removing the constants yields to

$$L(x;z,n) = z_1 \ln(1-\frac{2}{3}e^{-x}) - (n-z_1)x. \tag{3.13}$$

Note that

$$\frac{\mathrm{d}^2 L(x;z,n)}{\mathrm{d}x^2} = \frac{-z_1\frac{2}{3}e^{-x}}{(1-\frac{2}{3}e^{-x})^2} < 0, \tag{3.14}$$

so (3.13) is a concave function. To find the maximum of (3.13), I take its derivative and set it to zero. Let's name the solution as $\hat{d}$

$$\frac{\mathrm{d}L(x;z,n)}{\mathrm{d}x} = \frac{z_1\frac{2}{3}e^{-x}}{1-\frac{2}{3}e^{-x}} - (n-z_1) = 0. \tag{3.15}$$

Due to concavity of (3.13), and condisering (3.11) $d^*_{ML} = \max(0,\hat{d})$. The solution of (3.15) is $\hat{d} = -\ln(\frac{3}{2}(1-\frac{z_1}{n}))$. Note that $\hat{d} \geq 0$ if $\frac{3}{2}(1-\frac{z_1}{n}) \leq 1$ or $\frac{z_1}{n} \geq \frac{1}{3}$. So $d^*_{ML} = -\ln(\frac{3}{2}(1-\frac{z_1}{n}))$ if $\frac{z_1}{n} \geq \frac{1}{3}$, otherwise $d^*_{ML} = 0$.

Given n gene trees, to find the MAP estimate I need to solve

$$\begin{aligned} D^*_{MAP} &= \underset{x\geq 0}{\operatorname{argmax}} f_{D|Z_1}(x|z_1;n) \\ &= \underset{x\geq 0}{\operatorname{argmax}} P_{Z_1|D}(z_1|x) f_D(x) \end{aligned} \tag{3.16}$$

where $f_D(x)$ is a priori distribution for branch length.

The waiting time between the two consecutive speciation events is the length of the interior edge between them. Assuming that my species tree is a Yule tree, each species has a speciation rate λ, and this waiting time is exponentially distributed with mean $\frac{1}{2\lambda}$ [231]. So for the MAP estimate, I assume branch length D is exponentially distributed with rate 2λ. In this case the posteriori could be written as:

$$f_{D|Z_1}(x|z_1;n) \propto (1-\frac{2}{3}e^{-x})^{z_1}(\frac{1}{3}e^{-x})^{n-z_1}e^{-2\lambda x}. \tag{3.17}$$

Taking log of (3.17) and removing constants the function $L'(x)$ is defined as

$$L'(x) = z_1 \ln(1-\frac{2}{3}e^{-x}) - (n-z_1+2\lambda). \tag{3.18}$$

Note that

$$\frac{\mathrm{d}^2 L'(x;z,n)}{\mathrm{d}x^2} = \frac{-z_1\frac{2}{3}e^{-x}}{(1-\frac{2}{3}e^{-x})^2} < 0, \tag{3.19}$$

so (3.18) is a concave function. To find the maximum of (3.18), I take its derivative and set it to zero. Lets name the solution as $\hat{d'}$

$$\frac{\mathrm{d}L'(x)}{dx} = \frac{z_1\frac{2}{3}e^{-x}}{1-\frac{2}{3}e^{-x}} - (n-z_1+2\lambda) = 0. \tag{3.20}$$

Due to concavity of (3.18) and following (3.16) the MAP estimate is $d^*_{MAP} = \max(0, \hat{d'})$.

Solving (3.20) $\hat{d'} = -\ln\left(\frac{3}{2}(1-\frac{z_1}{n+2\lambda})\right)$. Note that $\hat{d'} \geq 0$ if $\frac{3}{2}(1-\frac{z_1}{n+2\lambda}) \leq 1$ or $\frac{z_1}{n+2\lambda} \geq \frac{1}{3}$. So, $d^*_{MAP} = -\ln\left(\frac{3}{2}(1-\frac{z_1}{n+2\lambda})\right)$ if $\frac{z_1}{n+2\lambda} \geq \frac{1}{3}$, otherwise the MAP estimate is $d^*_{MAP} = 0$. □

3.2.3 Calculation of quartet support

I now discuss how $\bar{z}$ defined in (3.1) can be efficiently computed. Note that in the worst case, there can be $\Theta(l^4)$ quartets around a single branch. Thus, simply enumerating all n_{ij} values and then getting the average can be very slow.

As noted before, each quartet around Q has each of its four leaves drawn from a different cluster of Q. Recall also that internal nodes of a tree produce a tripartition, and as [164] pointed out, any selection of two leaves from one side of the tripartition and one leaf from each remaining side gives a quartet tree mapped to that tripartition (Fig. 3.1). Let $\psi(Q)$ and $\psi(R)$ give the set of quartet trees around a quadripartition and mapped to a tripartition, respectively. Any quartet tree around Q that is induced by a gene tree will be mapped to two internal nodes in that gene tree (Fig. 3.1). Thus,

$$z_1 = \frac{1}{2m}\sum_{g=1}^{n}\sum_{u=1}^{l-2} |\psi(R_u^g) \cap \psi(Q)| \tag{3.21}$$

where R_u^g is a tripartition for node u in the gene tree g, and m is the number of quartets around Q.

I can efficiently compute the number of quartet topologies around Q that appear also in a tripartition R (i.e., $|\psi(R_u^g) \cap \psi(Q)|$) without computing $\psi(.)$ sets. I define a mapping between clusters around Q to clusters in R: map two sister clusters in Q to a cluster in R, and map the remaining two clusters in Q to the remaining two clusters in R (Fig. 3.1). For example, let $Q = AB|CD$ and $R = X|Y|Z$; a possible matching is to map A and B to Y, C to X, and D to Z. There are 12 such matchings between Q and R. For each matching, I can compute the number of quartet trees around Q that appear in R by multiplying the sizes of the intersection of pairs of clusters in Q and R that are mapped to each other. By enumerating all 12 matching and summing the resulting numbers, I get $|\psi(R_u^g) \cap \psi(Q)|$. Since finding the intersection of two clusters requires $\Theta(l)$, computing (3.21) would require $O(l^2 n)$ running time. However, I can do better. [165] have introduced a $\Theta(nl)$ algorithm to compute a sum similar to (3.21) for scoring a tripartition, based on a postorder traversal of gene trees (instead of analyzing each R_u^g separately). This algorithm can be adopted here to compute (3.21) in $\Theta(nl)$ (see Fig. 3.4 for the algorithm). Since scoring a tree requires scoring $l-3$ branches,

Theorem 9. Computing branch lengths and local posterior probabilities of a tree requires $\Theta(l^2 n)$ time.

3.2.4 Other considerations

I implemented my methods in ASTRAL, using the Colt [96] package for numerical computations. Handling missing data and unresolved gene trees requires extra care. When gene trees have missing data, to compute (3.21), instead of setting m to the number of quartets around Q, I need to set it to the average number of quartets present in gene trees (i.e., $\frac{1}{n}\sum_1^3 z_j$). Moreover, missing data can cause some genes to miss *all* quartets around Q; to account for this, I allow a different n for each branch, and set it to the number of genes that include at least one of the quartets around Q. To handle unresolved gene trees, similar to ASTRAL-II, I need to score the quadripartition against all $\binom{d}{3}$ tripartitions around a polytomy with degree d.

3.3 Materials and Methods

3.3.1 Datasets

I use both simulated and biological datasets.

```
**function** FREQ$(G, Q = (W,X)|(Y,Z))$
    $Q2 \leftarrow (W,Y)|(X,Z)$
    $Q3 \leftarrow (W,Z)|(X,Y)$
    $f1 \leftarrow F(G,Q)$
    $f2 \leftarrow F(G,Q2)$
    $f3 \leftarrow F(G,Q3)$
    $m \leftarrow (f1 + f2 + f3)/|G|$
    **return** $(f1/m, f2/m, f3/m)$
**function** F$(G, Q = (W,X)|(Y,Z))$
    $r \leftarrow 0$
    $S \leftarrow$ empty stack
    **for** $g \in G$ **do**
        **for** $u \in postOrder(g)$ **do**
            **if** $u$ is a leaf **then**
                $(w,x,y,z) \leftarrow (W[u], X[u], Y[u], Z[u])$
            **else**
                $(\mathbf{C}_{11}, \mathbf{C}_{12}, \mathbf{C}_{13}, \mathbf{C}_{14}) \leftarrow$ pull from $S$
                $(\mathbf{C}_{21}, \mathbf{C}_{22}, \mathbf{C}_{23}, \mathbf{C}_{24}) \leftarrow$ pull from $S$
                $(w,x,y,z) \leftarrow (\mathbf{C}_{11} + \mathbf{C}_{21}, \mathbf{C}_{12} + \mathbf{C}_{22}, \mathbf{C}_{13} + \mathbf{C}_{23}, \mathbf{C}_{14} + \mathbf{C}_{24})$
                $(\mathbf{C}_{31}, \mathbf{C}_{32}, \mathbf{C}_{33}, \mathbf{C}_{34}) \leftarrow (|W| - w, |X| - x, |Y| - y, |Z| - z)$
                $r \leftarrow r + I(\mathbf{C})$
            push $(w,x,y,z)$ to $S$
    **return** $r/2$
**function** I$(\mathbf{C})$
    **return**

        $\mathbf{C}_{10} \times \mathbf{C}_{21} \times \mathbf{C}_{32} \times \mathbf{C}_{33} + \mathbf{C}_{11} \times \mathbf{C}_{20} \times \mathbf{C}_{32} \times \mathbf{C}_{33} + \mathbf{C}_{12} \times \mathbf{C}_{23} \times \mathbf{C}_{30} \times \mathbf{C}_{31} +$
        $\mathbf{C}_{13} \times \mathbf{C}_{22} \times \mathbf{C}_{30} \times \mathbf{C}_{31} + \mathbf{C}_{30} \times \mathbf{C}_{21} \times \mathbf{C}_{12} \times \mathbf{C}_{13} + \mathbf{C}_{31} \times \mathbf{C}_{20} \times \mathbf{C}_{12} \times \mathbf{C}_{13} +$
        $\mathbf{C}_{32} \times \mathbf{C}_{23} \times \mathbf{C}_{10} \times \mathbf{C}_{11} + \mathbf{C}_{33} \times \mathbf{C}_{22} \times \mathbf{C}_{10} \times \mathbf{C}_{11} + \mathbf{C}_{10} \times \mathbf{C}_{31} \times \mathbf{C}_{22} \times \mathbf{C}_{23} +$
        $\mathbf{C}_{11} \times \mathbf{C}_{30} \times \mathbf{C}_{22} \times \mathbf{C}_{23} + \mathbf{C}_{12} \times \mathbf{C}_{33} \times \mathbf{C}_{20} \times \mathbf{C}_{21} + \mathbf{C}_{13} \times \mathbf{C}_{32} \times \mathbf{C}_{20} \times \mathbf{C}_{21}$
```

Figure 3.4: Frequency calculation algorithm. Input is a set of gene trees G and a quadripartition $Q = (W,X)|(Y,Z)$ where each cluster (e.g., X) is a bitset indexed by the species (thus, $X[u]$ is 1 if leaf u is in X and otherwise is 0). Output is the quartet frequency for each of the three topologies around that branch.

Table 3.1: Properties of simulated datasets. *l*: number of species; *n*: maximum number of genes (A-200 also has 50 and 200); *R*: number of replicates; *ILS*: average normalized RF [134] distance (AD) between the true species tree and true gene trees; *GE*: AD between true and estimated gene trees; *SE*: AD between true and ASTRAL species trees for A-200, with 50, 200, and 1000 genes, respectively).

	Cond.	l	n	R	ILS	GE	SE
A-200	Low-ILS	201	1000	100	15%	25%	6%,4%,3%
A-200	Med-ILS	201	1000	100	34%	31%	9%,6%,4%
A-200	High-ILS	201	1000	100	69%	47%	19%,10%,6%
Avian	1500bp	48	1000	20	47%	31%	5%
Avian	1000bp	48	1000	20	47%	39%	6%
Avian	500bp	48	1000	20	47%	54%	8%
Avian	250bp	48	1000	20	47%	67%	15%

Simulated data

I use two sets of simulated datasets from previous publications: the 200-taxon dataset (called A-200 here) from [165] and an avian dataset with 48 taxa from [161]. A-200 enables us to test accuracy under heterogeneous conditions with many species, and the avian dataset is used to compare local posterior against MLBS. For both datasets, gene trees are simulated using the MSC, and their branch lengths are then adjusted to be in substitution units and to deviate from the strict molecular clock. Sequence data are next simulated on the modified gene trees using GTR+Γ, and ML gene trees are estimated from the data. On the avian dataset, bootstrapped gene trees are also available. For both datasets, in addition to true species trees, I have estimated species trees (ASTRAL and NJst on estimated gene trees, and concatenation using ML). I show results for ASTRAL and true species tree here and show the rest in the appendix.

A-200: The 201-taxon datasets (200 ingroups plus an outgroup, treated like other taxa here) are simulated using SimPhy [149], and has three levels of ILS (Table 3.1), with true discordance that ranges from very low to very high (Fig. 3.5). Each replicate of the simulation has its own species tree, and the ILS level is controlled by changing the tree length (500k, 2M, and 10M generations). [165] generated species trees using the Yule process with two speciation rates (10^{-6} and 10^{-7} per generation) for each tree length, but here I combine the two rates into one dataset to get twice the number of replicates. SimPhy automatically introduces deviation from the strict clock by drawing species, gene, and gene/species-specific rate multipliers from predefined distributions. Similarly, the number of sites for each gene is randomly chosen. See [165] for details. ML Gene trees are estimated using FastTree-II [188], with a wide range of estimation error (Table 3.1 and Fig. 3.5). Species trees are estimated based on all 1000 genes per replicate, or on subsets of 200 or 50 genes. The ASTRAL species trees error for various datasets ranges between average 3% and 19% (Table 3.1).

Avian: The avian dataset has 48 taxa, and is simulated to emulate the whole-genome dataset of [105], possibly overestimating the true amount of ILS [161, 75]. Here, I use four conditions, with 20 replicates that each includes 1000 genes, all simulated based on the same avian-like species tree. My four conditions differ in terms of the number of sites per gene (250bp, 500bp, 1000bp, or 1500bp), creating varying levels of gene tree estimation error (Table 3.1). ML Gene trees are estimated using RAxML [232], and 200 replicates of bootstrapping are performed. Average ASTRAL species tree error ranges from 5% to 15%, depending on the gene tree error (Table 3.1). I used site-only MLBS to get BS support values. A single branch in my true tree was extremely short (almost a polytomy, with a length of 10^{-6}). When discussing branch length accuracy, I ignore that branch; results including that branch will also be shown for completeness.

Biological dataset:

I reanalyze four published datasets: a 103-taxon 424 gene plant dataset by [264], a 46-taxon 310 gene angiosperm dataset by [273], a 48-taxon 2022 (binned) supergene tree dataset by [105], and a 201-taxon 256 gene avian dataset by [189].

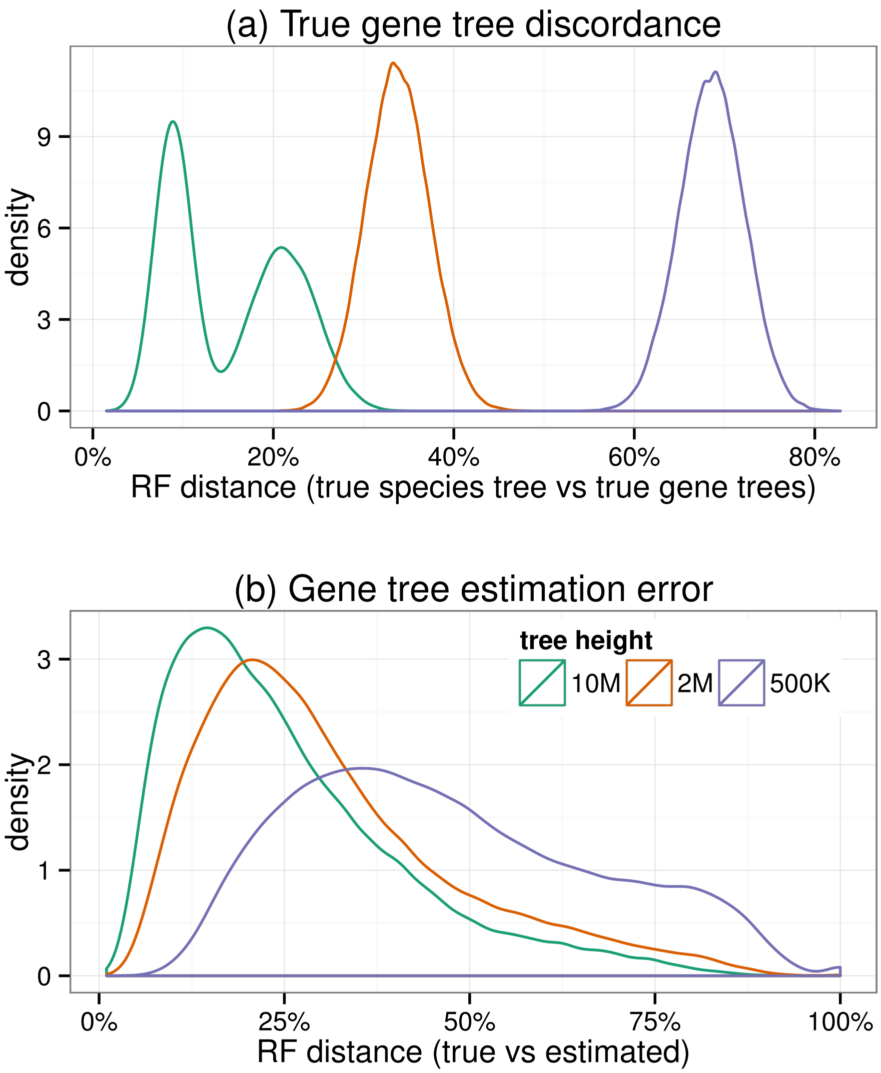


Figure 3.5: Properties of the A-200 dataset. (a) True gene tree discordance is shown, measured as the RF distance [196] between the true species tree and the true gene trees; Three levels of ILS are created by changing tree length (10M, 2M, or 500K generations), resulting in low, medium, and high discordance. The double pick is due to the fact that in [165] two different speciation rates are used, but here, I combine both speciation rates into one larger dataset. Bottom: Gene tree estimation error, measured as the RF distance between the true gene tree and the estimated gene tree. Note that the datasets with higher ILS also tend to have higher gene tree error.

3.3.2 Evaluation procedure

I study three questions in my evaluation:

- How accurate are branch lengths and support values when assumptions of my model are met?
- How do violations of the model assumptions impact the result?
- How do local posterior probabilities compare to site-only MLBS?

To answer these questions, I use both true gene trees and estimated gene trees to score both true and estimated species trees. For every internal branch in the species tree being scored, I also score its two alternative topologies. For example, if branch $AB|CD$ (Fig. 3.1) appears in the species tree, I also score $AD|BC$ and $AC|BD$. In my estimations, I use the Yule prior with fixed $\lambda = \frac{1}{2}$, but note that the true λ in my A-200 simulations ranges from 0.06 to 1.19.

With true species trees and true gene trees, all model assumptions are met. When estimated gene trees are used instead of true gene trees, I violate the assumption that input gene trees follow properties of the MSC model. When estimated species trees are scored, the locality assumption is potentially violated (i.e., each of the four clusters around a branch may be incorrect).

Measurement

Posterior: Despite the long-standing debate about correct interpretations of various measures of support [207, 69, 93, 243], biologists typically use support to judge branch reliability. A common practice is to ignore branches below a certain threshold of support and only interpret the remaining branches as biologically meaningful (0.95 for posterior and 70% for bootstrap are often used). My evaluation procedure takes a similar approach; I use varying thresholds of support and count the number of true and false branches with support at least equal to the threshold. For a threshold s, the measures I use are precision (the percentage of branches with support $\geq s$ that are correct), recall (the percentage of all true branches that have support $\geq s$), and false positive rate (FPR) (the percentage of all false branches that have support $\geq s$). I also draw the ROC curve (i.e., recall versus FPR).

MLBS and posteriors values are not directly comparable. Therefore, it is pointless to compare the precision or recall of MLBS and posterior for a given threshold. Instead, I use the ROC curve, which is agnostic to the exact interpretation of the threshold; it simply shows which method results in a better trade-off between false negative and false positive branches. Moreover, comparing to MLBS was only feasible on the avian dataset, where gene bootstrapping was doable. On the A-200 dataset (300 replicates each with 1000 genes of 201 taxa) bootstrapping was not computationally feasible.

Branch length: I measure branch length accuracy by comparing true and estimated lengths for each branch. Since this can be done only for correct branches, I measure branch length accuracy for the true species tree topology. Given b branches, and letting w_i and $\hat{w}_i$ indicate the true and estimated branch lengths, I use the logarithmic (log) error defined as $\frac{1}{b}\sum_1^b |\log_{10}(w_i) - \log_{10}(\hat{w}_i)|$. I also plot log of estimated versus true values. In addition, I show the root mean squared error, defined as $\sqrt{\frac{1}{b}\sum_1^b (w_i - \hat{w}_i)^2|}$. On Avian datasets, I compare the error of MP-EST and ASTRAL.

3.4 Results

3.4.1 A-200 dataset

Posterior

True trees: When true species trees are scored with true gene trees, the precision of

Table 3.2: Precision (and recall) of local posterior probabilities on A-200 dataset. For local posterior probability thresholds 0.95 and 0.99, I show the precision and recall (shown parenthetically) when true or ASTRAL species trees are scored with true or estimated gene trees.

		True species tree				ASTRAL species tree			
		True gene tree		Estimated gene tree		True gene tree		Estimated gene tree	
	n	.99	.95	.99	.95	.99	.95	.99	.95
Low ILS	1000	100.0(98.3)	100.0(98.7)	98.6(94.6)	98.4(95.4)	98.8(98.4)	98.8(98.8)	98.0(95.1)	97.8(95.9)
Low ILS	200	100.0(95.9)	100.0(96.8)	99.1(90.2)	98.9(91.9)	98.9(96.2)	98.8(97.1)	98.7(90.6)	98.5(92.4)
Low ILS	50	100.0(91.1)	100.0(93.2)	99.6(81.0)	99.3(85.0)	98.7(91.6)	98.6(93.6)	99.4(81.6)	99.0(85.6)
Med ILS	1000	100.0(95.1)	100.0(96.2)	98.9(90.8)	98.7(92.4)	99.3(95.4)	99.2(96.6)	98.6(91.2)	98.4(92.8)
Med ILS	200	100.0(89.2)	100.0(91.3)	99.4(82.6)	99.2(85.7)	99.4(90.0)	99.4(92.1)	99.3(83.3)	99.0(86.5)
Med ILS	50	100.0(79.0)	99.9(83.2)	99.7(70.0)	99.5(75.2)	99.4(81.0)	99.2(85.2)	99.7(71.3)	99.5(76.6)
High ILS	1000	100.0(83.0)	100.0(86.0)	99.4(77.5)	99.2(81.3)	99.7(84.2)	99.6(87.1)	99.4(78.3)	99.2(82.1)
High ILS	200	100.0(67.3)	100.0(72.8)	99.8(60.3)	99.6(66.6)	99.8(69.6)	99.7(75.1)	99.8(61.9)	99.6(68.3)
High ILS	50	100.0(46.0)	99.8(55.0)	99.8(38.6)	99.6(47.7)	99.8(50.0)	99.4(59.7)	99.8(41.1)	99.6(50.6)

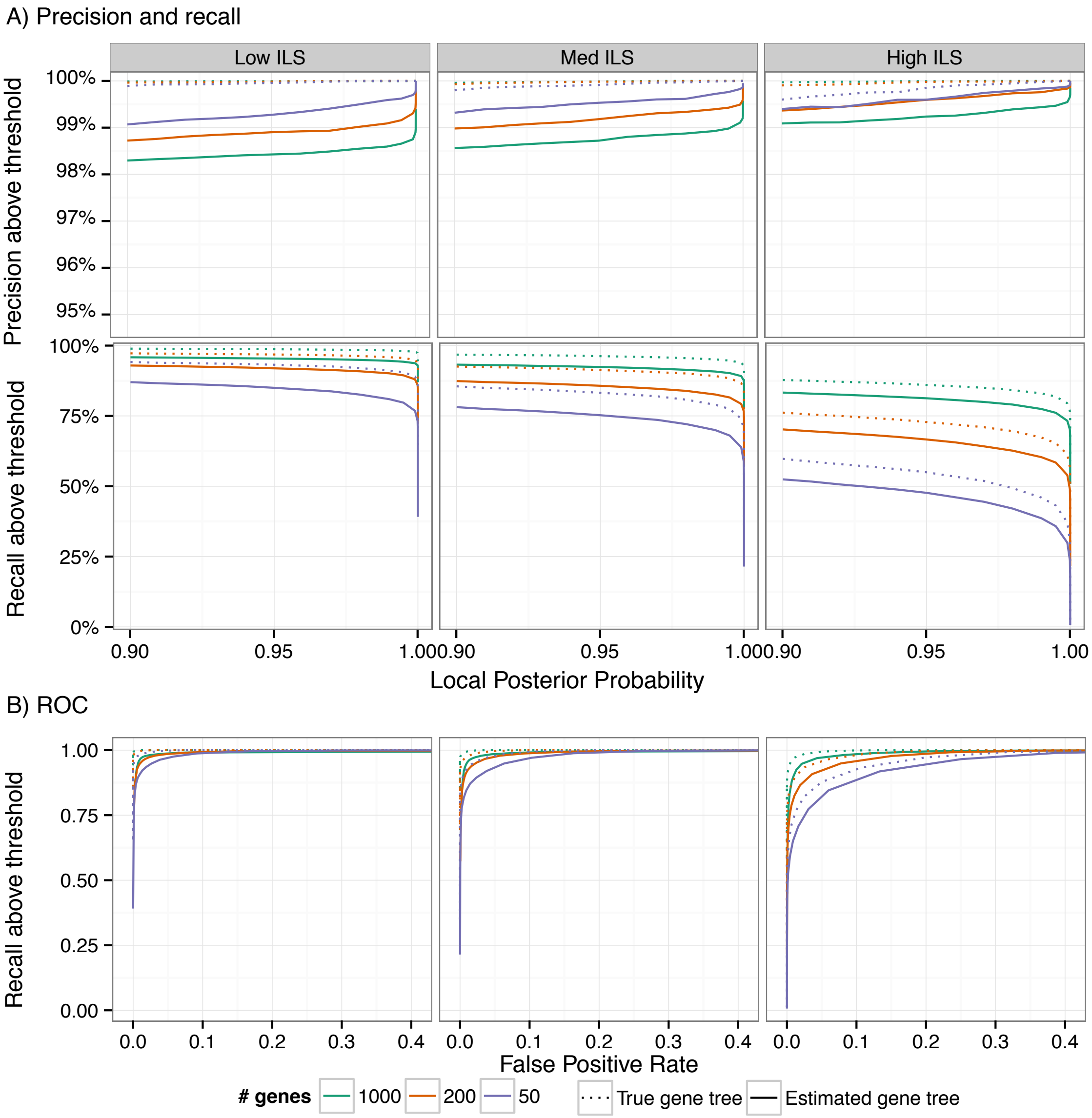


Figure 3.6: Accuracy of local posterior probability on the A-200 dataset with true species trees. A) the precision and recall of branches with local posterior probability above a threshold ranging from 0.9 to 1.0 using estimated gene trees (solid) or true gene trees (dotted). B) ROC curve (recall versus false positive rate) for thresholds ranging from 0 to 1 (figure trimmed at 0.4 fpr). Columns show different levels of ILS (each with 100 replicates). For each branch in each true species tree, all three alternatives are included in the analysis (one correct branch and two wrong branches). Thus, a total of 198*3*100=59400 branches are included for each model condition.

branches with 0.99 pp or higher is 100% for all model conditions, and is at least 99.8% for the 0.95 threshold (Table 3.2). Thus, there are very few false positive branches that have high local pp, a trend that continues if I further lower the threshold to 0.9 (Fig. 3.6). With the 0.95 threshold, the percentage of true branches that are recovered (recall) ranges from very high (98.7%) for the model condition with low ILS and 1000 gene trees to moderate (55.0%) for the most challenging dataset with high ILS and only 50 genes (Table 3.2). As desired, increasing ILS and reducing the number of genes both reduce the recall while maintaining high precision (Fig. 3.6).

Estimated gene trees: When true species trees are scored on estimated gene trees instead of true gene trees, precision slightly drops from 100% to between 98.4% and 99.8%. The recall, however, is impacted more and is reduced by as much as 10% (Table 3.2 and Fig. 3.6). Thus, gene tree estimation error has a small impact on the precision but a substantial impact on the recall.

The impact of the threshold is also interesting. Going from the 0.99 to 0.95 threshold, as expected, recall improves (e.g., from 39% to 48% for high ILS, 50 genes) but reductions in the precision are very small (at most 0.3%). Thus, a 95% threshold results in meaningful improvements in the recall without substantially sacrificing the precision. The ROC curves (Figs. 3.7B and 3.6) further explore the tradeoff between increasing recall and allowing false positive branches.

Estimated species trees: By scoring estimated species trees I study the impact of violating the locality assumption. I show results for ASTRAL here, but similar results are obtained with NJst and concatenation (Figs. 3.6 and 3.9).

Precision and recall are remarkably similar between ASTRAL and true species trees, especially with estimated gene trees (Table 3.2). Comparing true species trees and ASTRAL on estimated gene trees, the precision is reduced at most by 0.6% while the recall is surprisingly increased, by up to 2.9%. The impact of violating locality assumptions is more

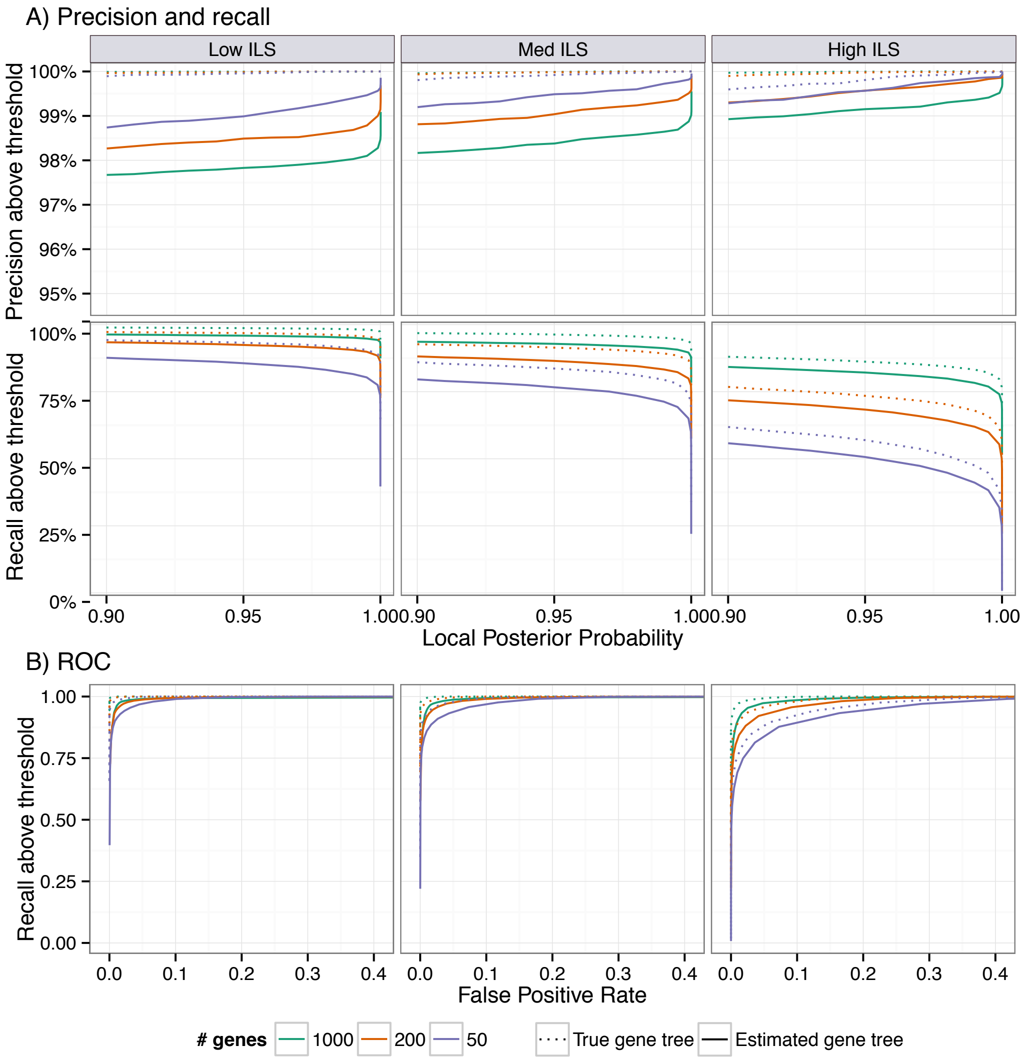


Figure 3.7: Evaluation of local posterior probability on the A-200 dataset with ASTRAL species trees. See Figures 3.6-3.9 for other species trees. A) Precision and recall of branches with local posterior probability above a threshold ranging from 0.9 to 1.0 using estimated gene trees (solid) or true gene trees (dotted). B) ROC curve (recall versus false positive rate) for varying thresholds (figure trimmed at 0.4 FPR). Columns show different levels of ILS.

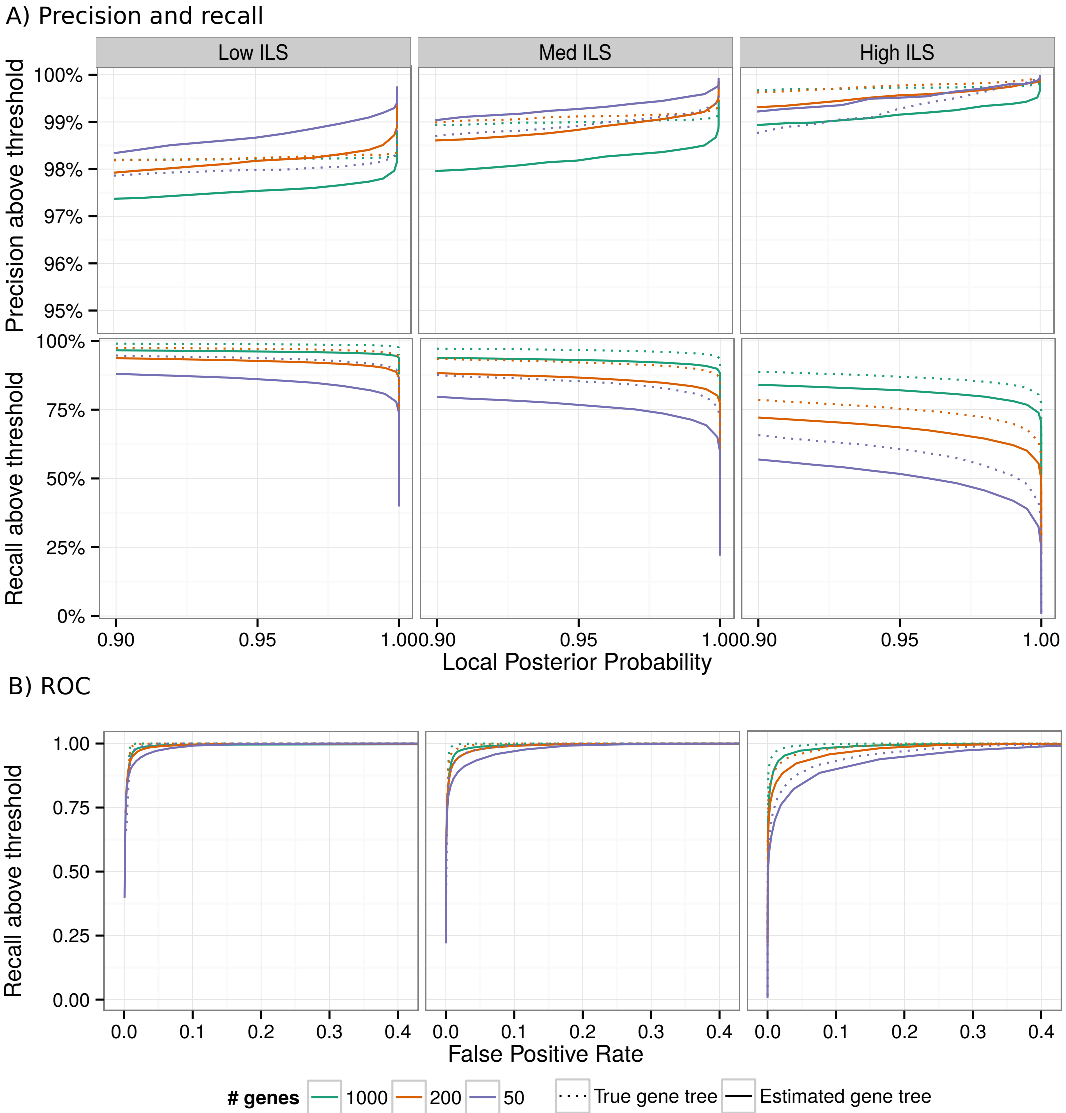


Figure 3.8: Accuracy of local posterior probability on the A-200 dataset with NJST species trees. A) the precision and recall of branches with local posterior probability above a threshold ranging from 0.9 to 1.0 using estimated gene trees (solid) or true gene trees (dotted). B) ROC curve (recall versus false positive rate) for thresholds ranging from 0 to 1 (figure trimmed at 0.4 fpr). Columns show different levels of ILS (each with 100 replicates). For each branch in each true species tree, all three alternatives are included in the analysis (one correct branch and two wrong branches). Thus, a total of 198*3*100=59400 branches are included for each model condition.

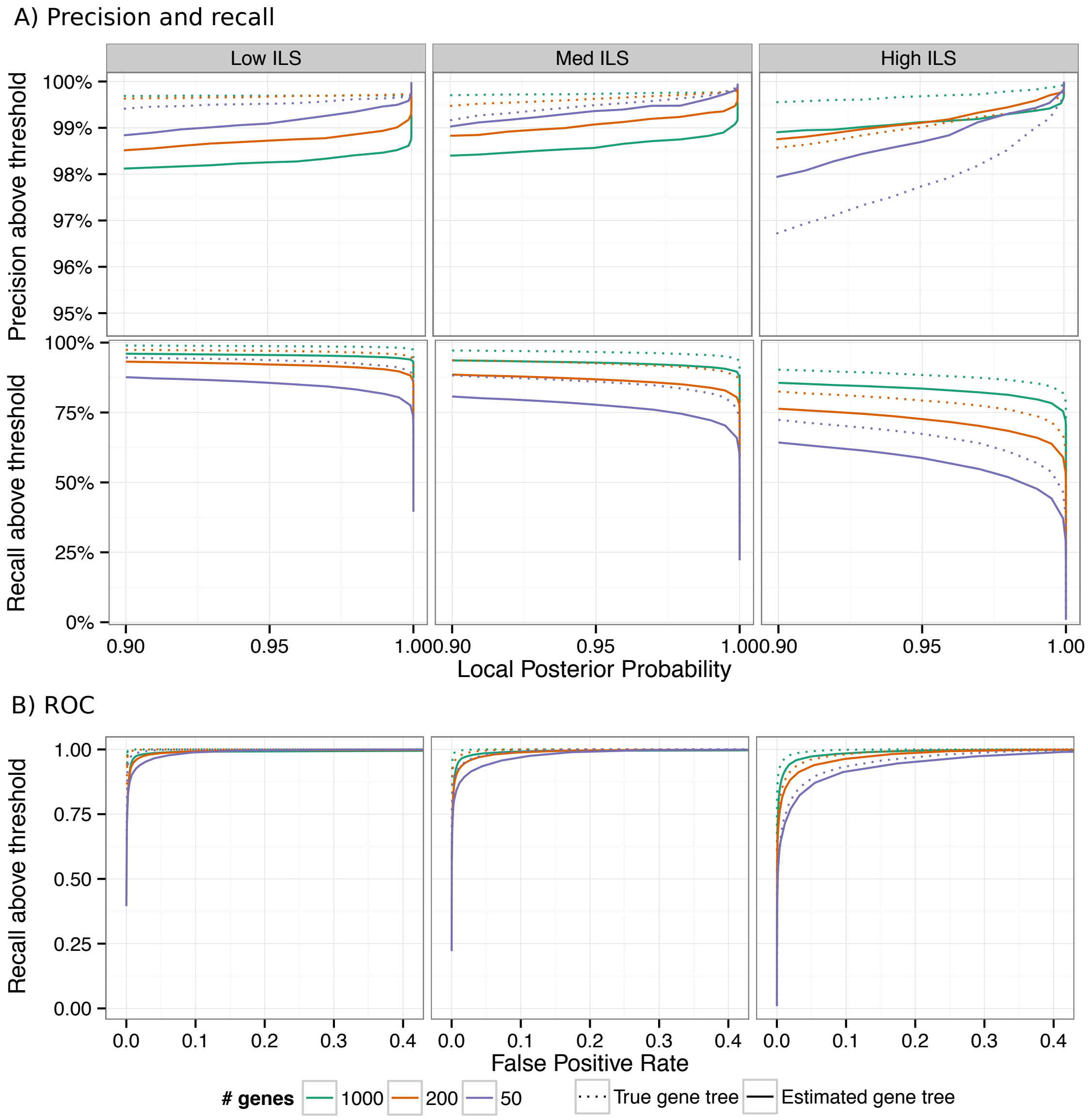


Figure 3.9: Accuracy of local posterior probability on the A-200 dataset with concatenation. A) the precision and recall of branches with local posterior probability above a threshold ranging from 0.9 to 1.0 using estimated gene trees (solid) or true gene trees (dotted). B) ROC curve (recall versus false positive rate) for thresholds ranging from 0 to 1 (figure trimmed at 0.4 fpr). Columns show different levels of ILS (each with 100 replicates). For each branch in each true species tree, all three alternatives are included in the analysis (one correct branch and two wrong branches). Thus, a total of 198*3*100=59400 branches are included for each model condition.

Table 3.3: Branch length accuracy on the A-200 dataset. Logarithmic error (Log Err) and root mean squared error (RMSE) are shown for true species trees scored with true gene trees or estimated gene trees (Est. gt).

		Log Err		RMSE	
Dataset	*n*	*True gt*	*Est. gt*	*True gt*	*Est. gt*
Low ILS	1000	0.10	0.42	5.57	6.75
Low ILS	200	0.16	0.44	6.22	6.99
Low ILS	50	0.25	0.48	6.84	7.29
Med ILS	1000	0.03	0.20	0.22	0.86
Med ILS	200	0.07	0.22	0.44	0.91
Med ILS	50	0.13	0.26	0.74	1.05
High ILS	1000	0.06	0.15	0.03	0.13
High ILS	200	0.11	0.18	0.07	0.15
High ILS	50	0.18	0.24	0.14	0.19

pronounced when true gene trees are used. Once again, precision is reduced (by as much as 1.4%) and the recall is increased (up to 4.7%). Thus, moderate violations of the locality assumption have minimal impact.

I note that in my analyses, deviations from the locality assumption are moderate but realistic, as most ASTRAL trees have a relatively high accuracy (Table 3.1). The least accurate ASTRAL trees have 19% RF distance to the true species tree (50 genes and high ILS), which means 81% of the clusters in the estimated tree remain correct. Nevertheless, it is interesting that violating the locality assumption for up to 19% of clusters has minimal impact on the precision and positive impact on the recall.

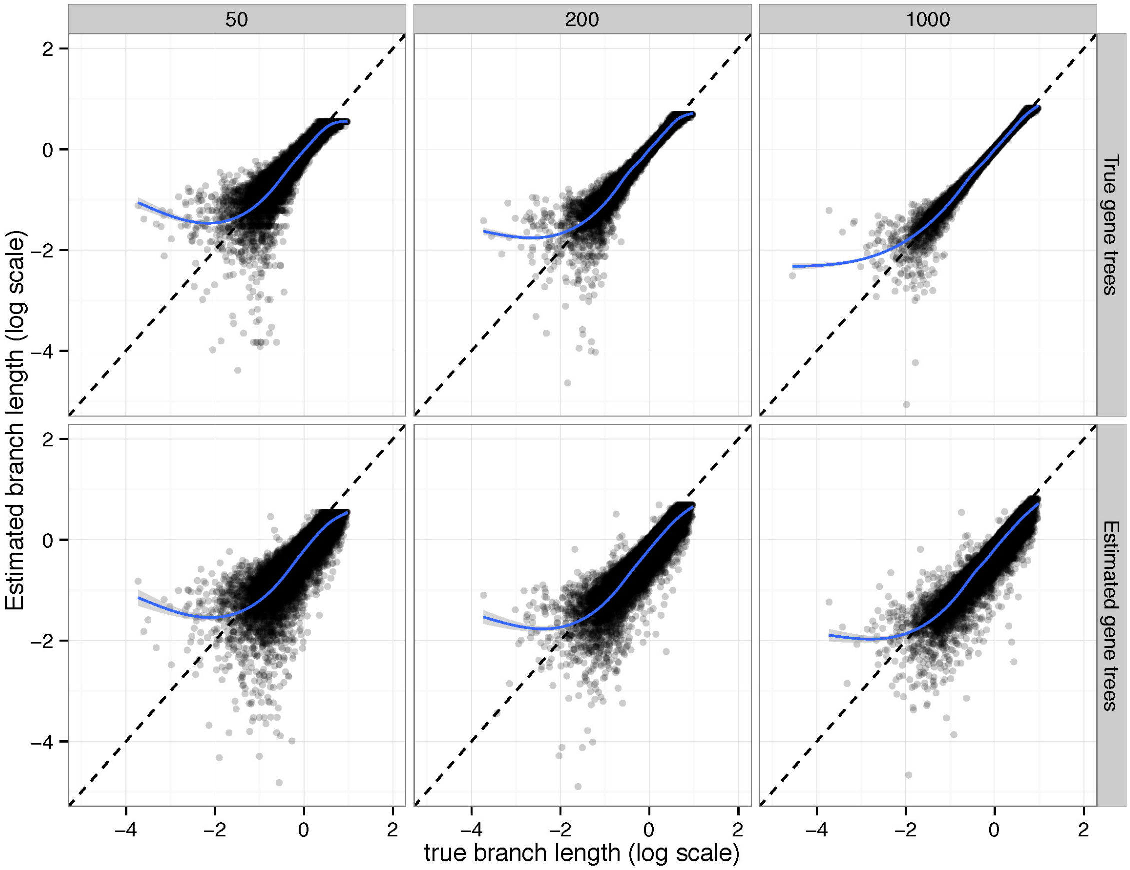


Figure 3.10: Branch length accuracy on the A-200 dataset with Medium ILS. See Figs. 3.12 and 3.11 for low and high ILS. The estimated branch length is plotted against the true branch length in log scale (base 10). Blue line: a fitted generalized additive model with smoothing [271].

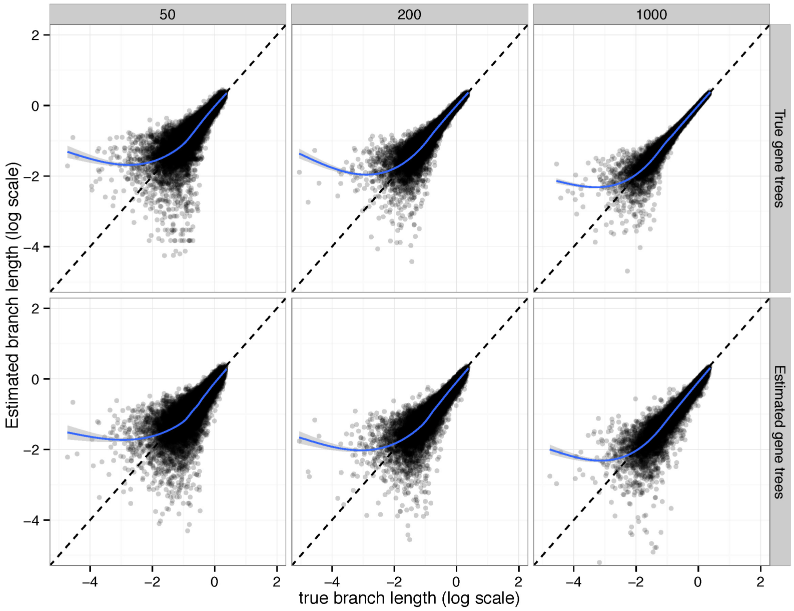


Figure 3.11: Branch length accuracy on the A-200 High ILS dataset.Estimated branch length against true branch length is plotted in log scale (base 10). Line: fitted generalized additive model with smoothing [271].

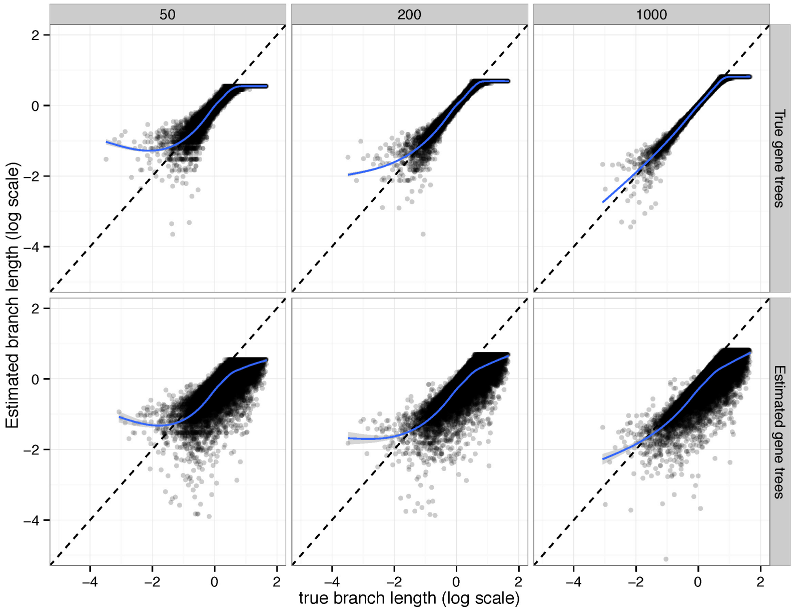


Figure 3.12: Branch length accuracy on the A-200 Low ILS dataset.Estimated branch length against true branch length is plotted in log scale (base 10). Line: fitted generalized additive model with smoothing [271].

Branch Length

The accuracy of branch lengths is dramatically impacted by gene tree estimation error, the number of genes, and the amount of ILS (Table 3.3). With 1000 true gene trees, the logarithmic error is very low, ranging from 0.03 to 0.10 (which correspond to branches that on average are respectively 7% or 25% shorter or longer than true branches). As the number of genes is reduced, the logarithmic error predictably goes up, but with true gene trees, it never exceeds 0.25. Moreover, with true gene trees, the error is largely unbiased, except perhaps for very short or long branches that are hard to estimate correctly with a limited number of genes (Figs. 3.10 and 3.12-3.11).

Branch length error dramatically increases when estimated gene trees are used. Low ILS conditions are impacted the most by gene tree error (Table 3.3 and Fig. 3.12). For example, with 1000 estimated gene trees and low ILS, log error is 0.42, corresponding to estimated branches that are on average 2.6 times too short or long. Moreover, the error is biased towards underestimation, especially for low ILS (Figs. 3.10, 3.12, and 3.11). This pattern is not surprising because as I will show, gene tree error tends to increase observed gene tree discordance and branch lengths are a function of observed discordance.

3.4.2 Avian

On the avian dataset, I compare local posterior probabilities against branch support generated using site-only MLBS with estimated gene trees and ASTRAL species trees. Here, I also study the impact of increasing levels of gene tree estimation error by decreasing the number of sites per gene from 1500bp to 250bp.

3.4.3 Posterior and MLBS

The precision of local posterior probability is 100% for the 0.99 threshold, regardless of the numbers of sites, but the recall ranges from 81% for the 1500bp model condition to 69% for 250bp (Table 3.4 and Figs. 3.13 and 3.14). Precision is at least 99.8% for the 0.95 threshold, and the recall is between 71.5% and 84.7%, depending on the model condition (an improvement of 2% to 5% compared to the 0.99 threshold). Lowering the support threshold all the way to 0.7 still retains at least 99.1% accuracy and increases the recall to between 78.3% and 91.4%. Therefore, the local posterior probabilities allow very few false positives with high support but also miss some true positives (and thus may be conservative).

Nevertheless, local posterior probabilities are less conservative than MLBS support values and have better recall. As the ROC curves show (Fig. 3.15), for the same number of false positives branches, local posterior probabilities result in better recall than MLBS. This pattern is more pronounced for shorter alignments, which have increased gene tree error. For example, for the 250bp model condition, if I choose a support threshold that results in 0.01 false positive rate, with local posterior values, I still recover 84% of correct branches, whereas with MLBS, the same false positive rate results in retaining 70% of correct branches. Thus, for a desired level of precision, better recall can be obtained using local posterior probabilities.

Branch Length

Branch length accuracy on the avian dataset was a function of gene tree estimation error whether ASTRAL or MP-EST was used (Table 3.5). With true gene trees, branch length log error was only 0.06, corresponding to branches that are about 14% shorter or longer than the true branch. As gene tree estimation error increases with reduced number of sites (see Table 3.1 for gene tree error statistics), the branch length error also increases. Thus, while 1500bp genes give 0.17 log error, 250bp genes result in 0.59 error, which corresponds

Table 3.4: Support accuracy on the avian dataset. Accuracy (and recall) of BS and local posterior probability is show for three thresholds: 0.7, 0.95, and 0.99.

	BS			Local PP with Estimated gene trees			Local PP with True gene trees		
sites	0.70	0.95	0.99	0.70	0.95	0.99	0.70	0.95	0.99
1500	99.4(93.5)	100.0(89.6)	100.0(85.9)	99.9(90.7)	100.0(84.7)	100.0(81.3)	99.9(93.6)	100.0(90.2)	100.0(86.2)
1000	99.6(89.0)	100.0(79.7)	100.0(74.3)	100.0(91.4)	100.0(84.1)	100.0(80.1)	100.0(94.5)	100.0(91.4)	100.0(87.7)
500	98.0(75.7)	99.8(69.8)	100.0(66.8)	99.5(87.7)	99.8(79.8)	100.0(74.4)	99.0(97.4)	99.4(93.8)	99.5(90.6)
250	95.9(71.7)	99.6(63.2)	100.0(57.4)	99.1(78.3)	100.0(71.5)	100.0(69.2)	97.1(97.7)	98.6(94.7)	99.2(92.6)

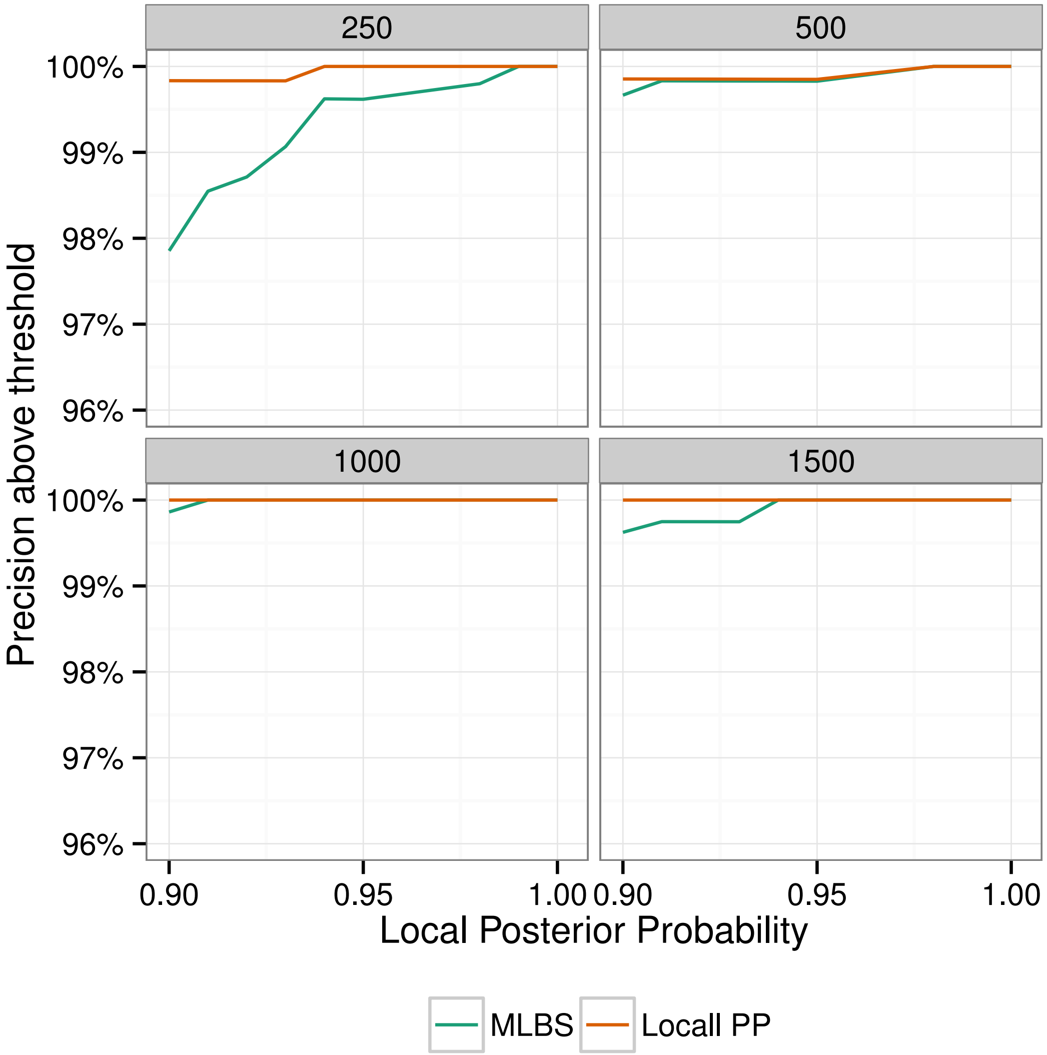


Figure 3.13: Precision of local posterior probability on the Avian dataset with ASTRAL. Precision of branches with local posterior probability above a threshold ranging from 0.9 to 1.0 based on MLBS and local posterior probability (PP) support values. Boxes show different numbers of sites per gene (controlling gene tree estimation error).

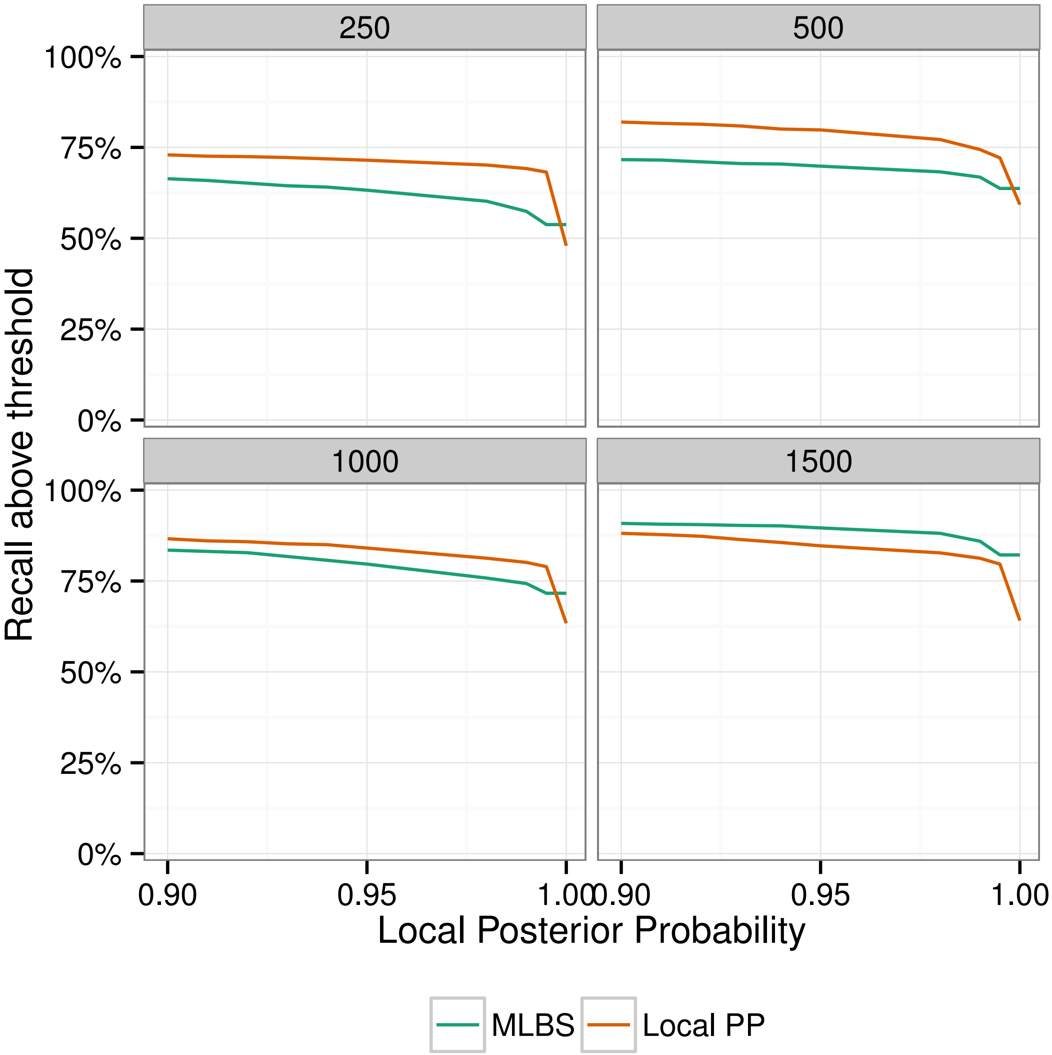


Figure 3.14: Recall of local posterior probability on the Avian dataset with ASTRAL. Recall of branches with local posterior probability above a threshold ranging from 0.9 to 1.0 based on MLBS and local posterior probability (PP) support values. Boxes show different numbers of sites per gene (controlling gene tree estimation error).

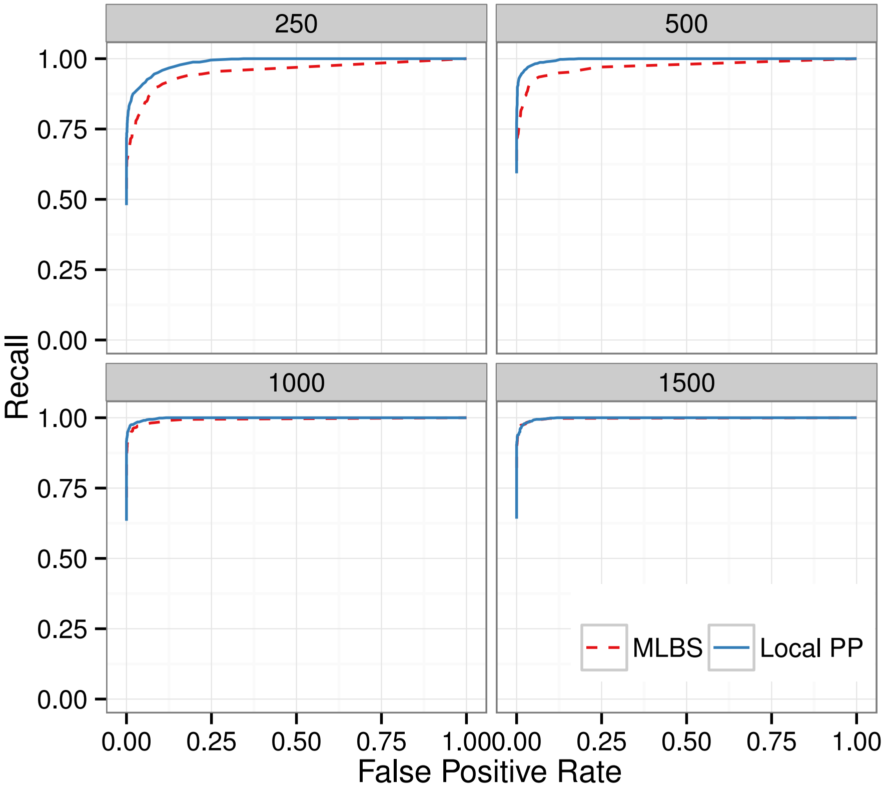


Figure 3.15: ROC curve for the avian dataset based on MLBS and local posterior probability (PP) support values. Boxes show different numbers of sites per gene (controlling gene tree estimation error).

Table 3.5: Branch length accuracy for the avian dataset. Logarithmic error and root mean squared error are shown for true species trees scored with true gene trees or estimated gene trees with various numbers of sites using ASTRAL and MP-EST. An extremely short branch with length 10^{-6} was removed from the calculations, but error including that branch is shown parenthetically.

	Log Err		RMSE	
# sites	ASTRAL	MPEST	ASTRAL	MPEST
True gt.	0.06 (0.10)	0.07 (0.11)	0.44 (0.44)	0.30 (0.30)
1500	0.17 (0.20)	0.14 (0.18)	0.83 (0.83)	0.70 (0.70)
1000	0.22 (0.27)	0.22 (0.25)	1.08 (1.07)	1.01 (1.00)
500	0.37 (0.42)	0.42 (0.46)	1.65 (1.64)	1.65 (1.64)
250	0.59 (0.63)	0.81 (0.84)	2.25 (2.24)	2.28 (2.26)

to branches that are on average 3.9 times too short or long. Moreover, unlike true gene trees, the error in branch lengths estimated based on estimated gene trees is biased toward underestimation (Fig. 3.16), a pattern that increases in intensity with shorter alignments.

ASTRAL and MP-EST have similar branch length accuracy measured by log error for highly accurate gene trees, but ASTRAL has an advantage with increased gene tree error (Fig. 3.18, Table 3.5). Error measured by RMSE (which emphasizes the accuracy of long branches) is comparable for the two methods, but MP-EST has a slight advantage given accurate gene trees.

3.4.4 Biological datasets

For each biological dataset, I show MLBS support and the local posterior probabilities, computed based on RAxML gene trees available from respective publications. I also collapse gene tree branches with less than 33% bootstrap support and use these collapsed gene trees to draw local posterior probabilities. For ease of discussion, I show local posterior probabilities as percentages and refer to them simply as posterior or collapsed posterior (for values based on collapsed gene trees). I discuss the confidence in important branches in each tree.

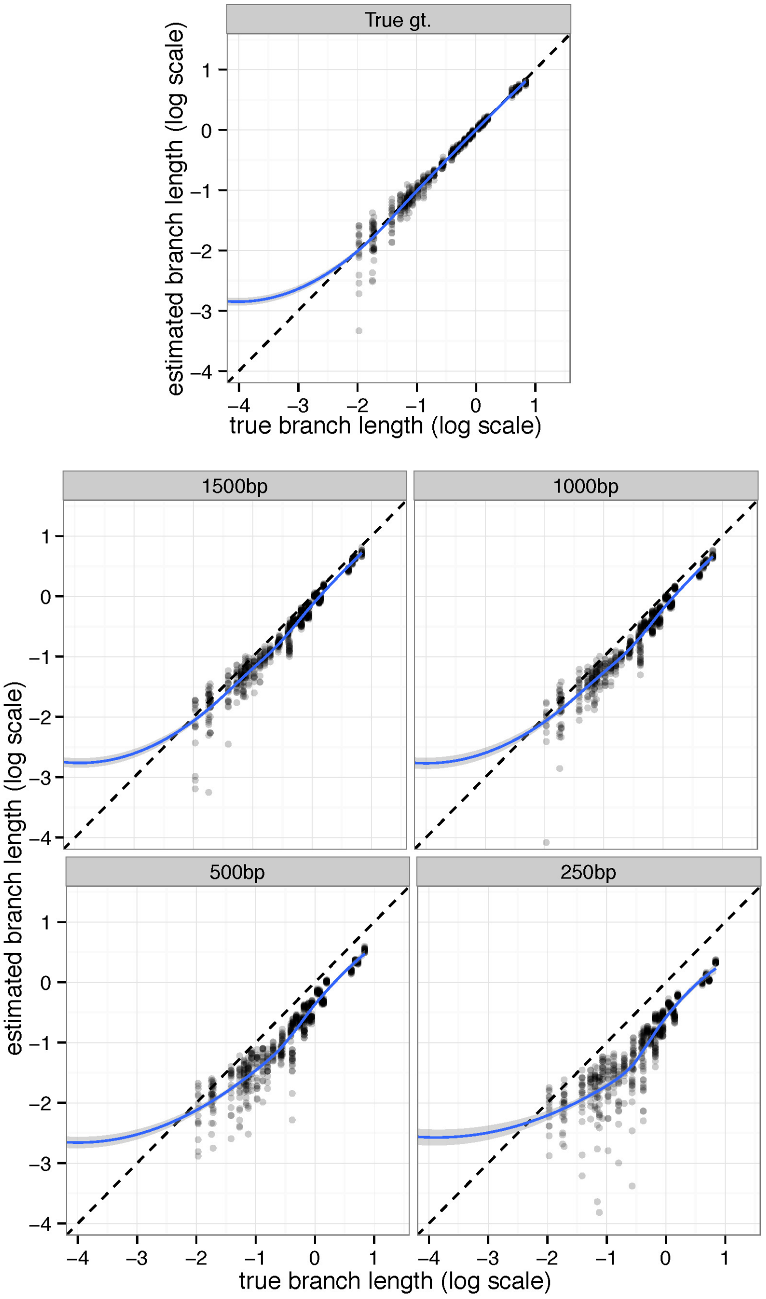


Figure 3.16: ASTRAL branch length accuracy on the avian dataset. Log transformed estimated branch lengths are shown versus true branch lengths, and a generalized additive model is fitted to the data. One branch with length 10^{-6} is trimmed out here, but full results, including MP-EST, is shown in Fig. 3.17.

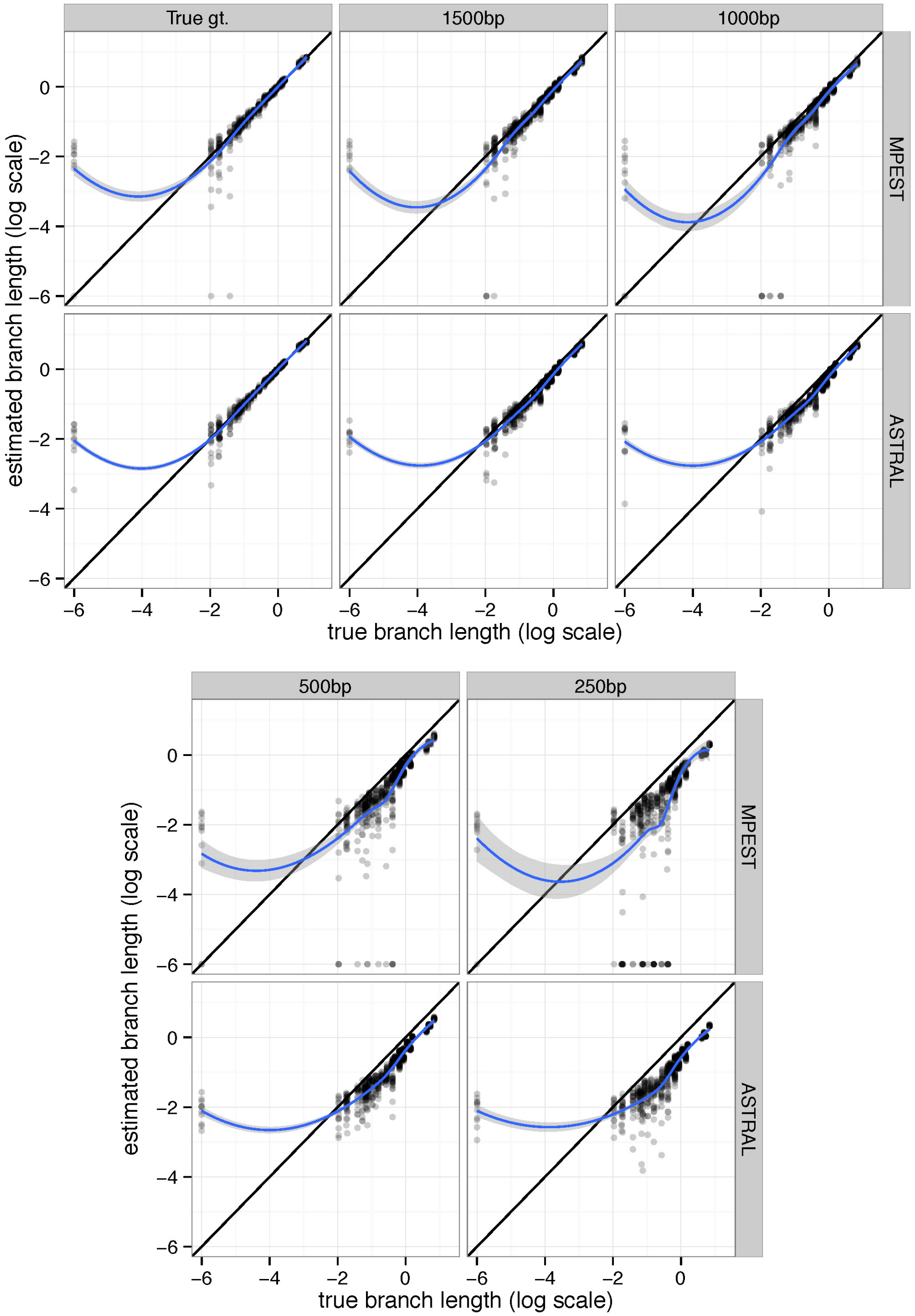


Figure 3.17: Branch length accuracy on the Avian dataset. Estimated branch length against true branch length is plotted in log scale (base 10). Line: fitted generalized additive model with smoothing [271].

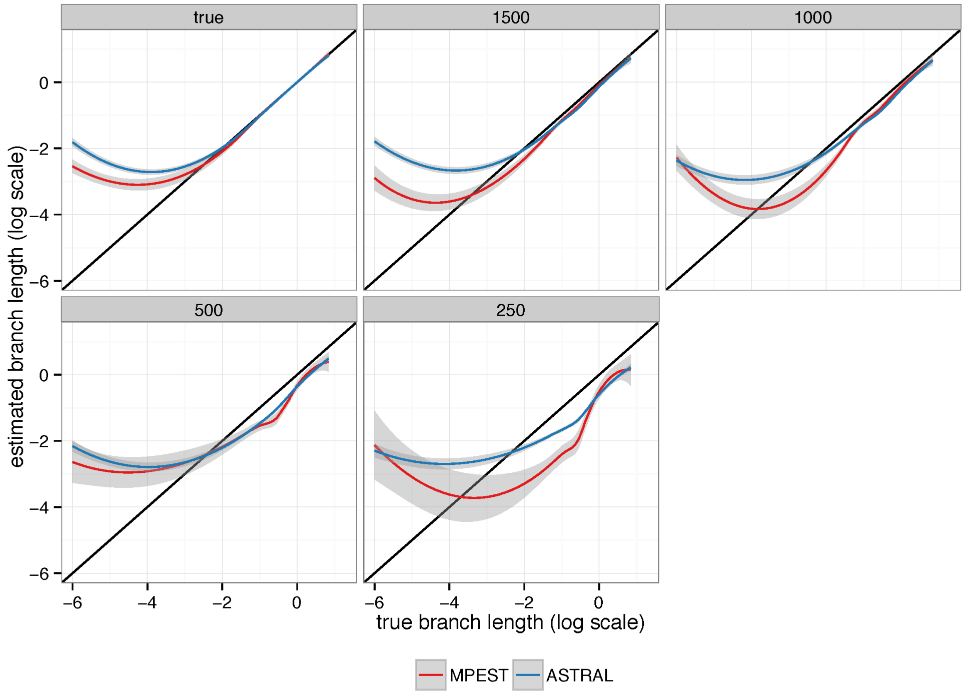


Figure 3.18: Summarized branch length accuracy on the Avian dataset. Estimated branch length using MPEST, and ASTRAL methods is plotted against true branch length in log scale (base 10). Lines show a fitted generalized additive model with smoothing [271].

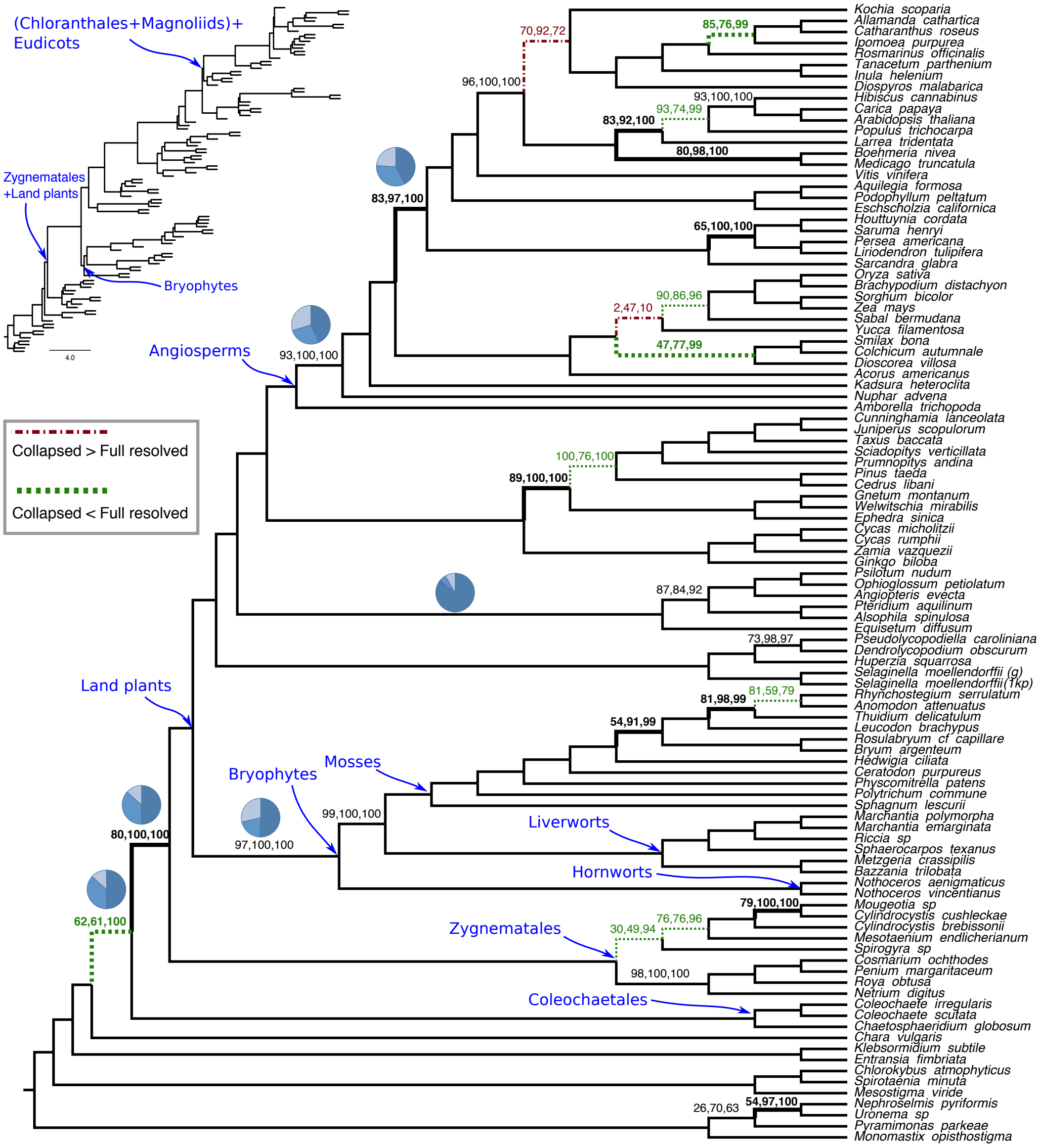


Figure 3.19: ASTRAL tree on the 1KP dataset of [264] (103 taxa and 424 genes). On each branch, three support values are shown: BS (using site-only MLBS), local posterior computed on fully resolved ML gene trees, and local posterior computed on collapsed ML gene trees (removing branches with < 33% BS). Branches with no designation have 100% support according to all three measures. Dotted/green lines (dashed/red lines): collapsing low support gene trees branches increases (decreases) posterior by at least 10%. Bold: collapsed posterior is at least 10% higher than BS. Inset: ASTRAL tree with branch lengths in coalescent units using collapsed genes (terminals lengths drawn arbitrarily). Pie-charts (for selected edges): relative frequencies of the three quartet topologies around a branch in collapsed gene trees.

1KP: Three of the key relationships studied by [264] are the sister branch to land plants, the base of the angiosperms, and the relationship among Bryophytes (hornworts, liverworts, and mosses). In the ASTRAL tree, many branches have full support regardless of the measure of the support used, but the remaining branches reveal interesting patterns (Fig. 3.19). The sister relationship between Zygnematales and land plants receives a moderate 80% BS, but has 100% posterior. [264] also recovered this relationship by concatenation of various data partitions. There are 12 other branches that have collapsed posteriors that are at least 10% higher than BS (Fig. 3.19); no branch has substantially higher BS than collapsed posterior. Collapsed posterior for monophyly of Bryophytes and for Amborella as sister to other angiosperms are 100% (compared to 97% and 93% BS, respectively).

When I collapse low support branches in gene trees, posterior goes up for several branches: nine branches have improvements of 10% or more, and only two branches have comparable reductions. An interesting case is Coleochaetales as sister to Zygnematales+land plants, which has only 62% BS and 61% posterior, but has 100% collapsed posterior. Finally, note that several branches have low posterior, even after collapsing.

My estimated branch lengths are short for several nodes. For example, the branch that unites (Chloranthales+Magnoliids) and Eudicots has a length of 0.14 in coalescent units. Other branches that have been historically hard to recover also tend to have short branches; however, these are not necessarily extremely short branches that would implicate anomaly zone (two adjacent branches below 0.1 will result in an anomaly zone [202]). For example, Bryophytes had a length of 0.29, and Zygnematales+land plants had a length of 0.28. These values, while short, are not below the often-cited 0.1 threshold. Moreover, as my simulation study showed, I caution that branch lengths tend to be underestimated because of gene tree error and these numbers should be treated as lower bounds.

Angiosperms: [273] have used 310 genes to study the base of the angiosperm tree, a question of intense debate [226, 279, 82, 220]. Unlike the MP-EST tree by [273], but similar

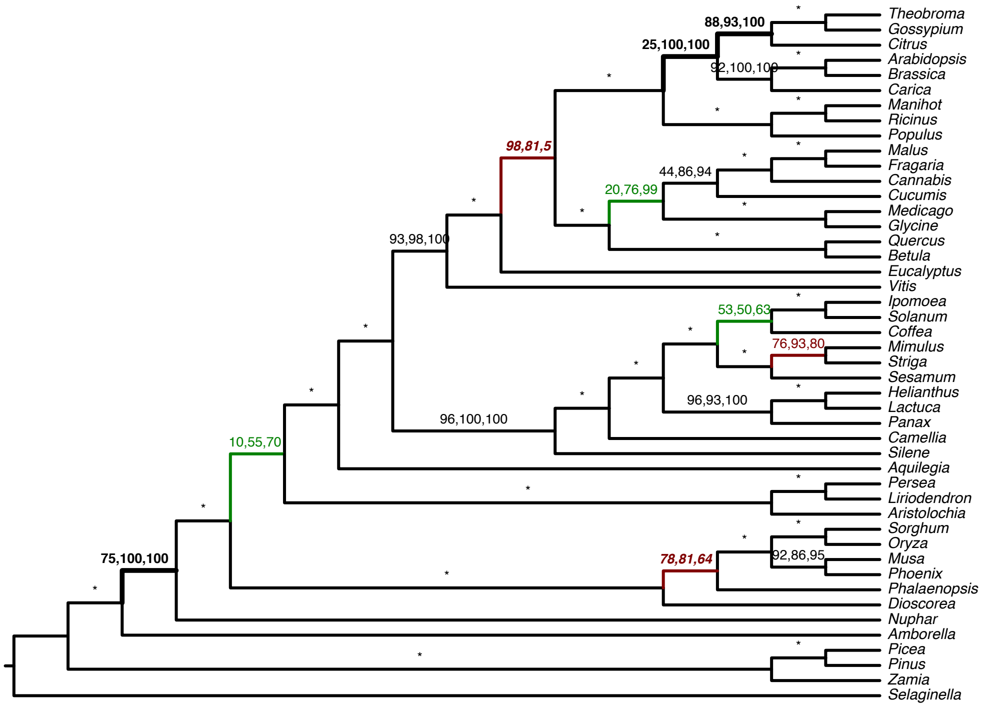


Figure 3.20: Support on the angiosperm dataset. Three forms of support values are shown for the angiosperm dataset of Xi et al [273]. Evaluation of local posterior probability on the A-200 dataset with secies trees. See Figures 3.6-3.9 for other species trees. A) Precision and recall of branches with local posterior probability above a threshold ranging from 0.9 to 1.0 using estimated gene trees (solid) or true gene trees (dotted). B) ROC curve (recall versus false positive rate) for varying thresholds (figure trimmed at 0.4 FPR). Columns show different levels of ILS

to concatenation on this dataset and ASTRAL and concatenation on 1KP, the ASTRAL tree recovers Amborella as sister to the rest of the angiosperms. This relationship has 75% BS, but its posterior and collapsed posterior are 100% (Fig. 3.20). The length of this branch is estimated to be 0.160, almost exactly matching the length estimated on the 1KP dataset (0.156).

Avian (genomes): [105] used whole-genomes of 48 bird species to resolve long-standing questions about relationships at the base of Neoaves. I reanalyzed their 2022 supergene trees (binned gene trees; see [161]) using ASTRAL, which produced a tree with a wall of short branches at the base of Neoaves (Fig. 3.21); 12 branches are below 0.1 coalescent

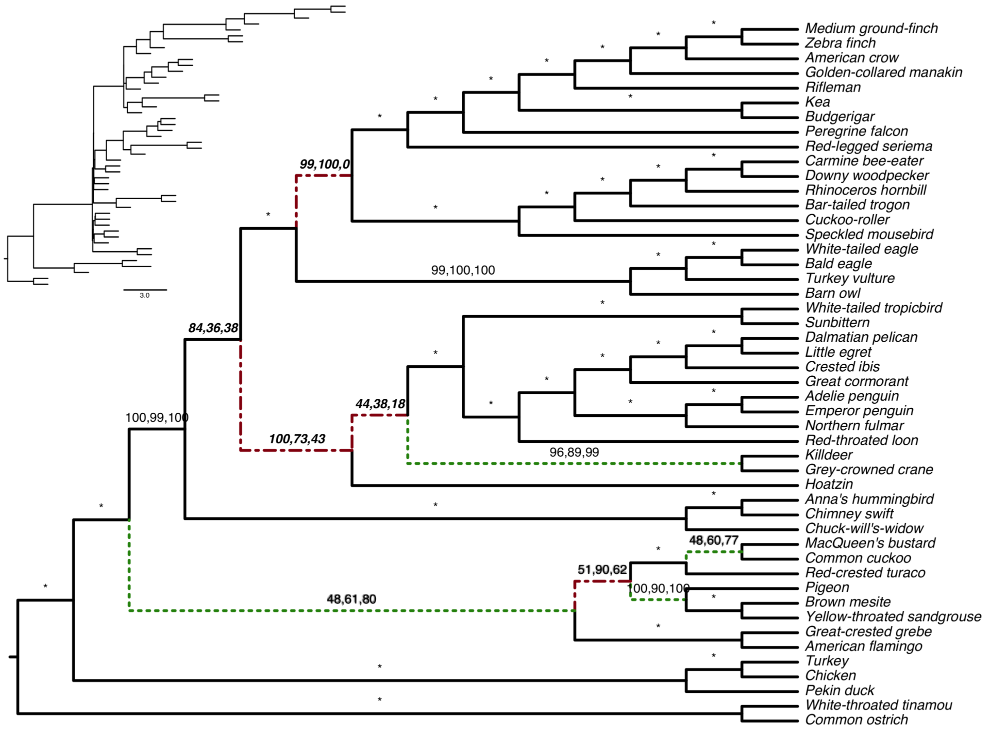


Figure 3.21: Support on the avian dataset. Three forms of support values are shown for the avian dataset of Jarvis et al [105]. MLBS support (site-only), local posterior computed on fully ML resolved gene trees, and local posterior computed on ML gene trees with branches with less than 33% support collapsed. Branches marked with an asterisk (*) have 100% support with all three measures. Bold: local support on collapsed gene trees is at least 10% higher than MLBS support. Bold italic: local support on collapsed gene trees is at least 10% lower than MLBS support. Dotted/green lines (dashed/red lines): collapsing low support branches in gene trees increases (decreases) local posterior probability by at least 10%. Inset: branch lengths in coalescent units. The length of terminal branches are drawn arbitrarily.

units, and another 11 are below 0.5. However, when low support branches in gene trees were collapsed, branch lengths increase by a median of 0.23 units. Nevertheless, 11 branches remain below 0.1, and another four branches are below 0.5. My results support the hypothesis that avian big bang [66] gave rise to very short branches, but quantifying the exact lengths remains difficult because of gene tree error.

Support values on the avian tree also revealed interesting patterns. Despite the large number of supergene trees (2022), which should increase support (Fig. 3.3), several key branches have low posterior. For example, the position of Hoatzin (arguably, the most difficult avian order to place) and the two branches around it have collapsed posterior below 50% and posterior below 75% (Fig. 3.21). Similarly, at the base of Neoaves, a clade containing land birds, water birds, and Caprimulgiformes has full support, but the sister to this large group has only 61% posterior and 80% collapsed posterior. However, other challenging relationships have full support (e.g., falcons as sister to parrots+passerines, and seriema as sister to this group, or a clade containing owls, eagles, and vultures). Thus, despite large number of input trees, posterior probabilities reveal some uncertainty. Finally, on this dataset, unlike the 1KP dataset, four branches have substantially lower posterior compared to BS, and posteriors are higher than collapsed posteriors in some cases.

Avian (high sampling): [189] used their dataset of 259 genes and 201 species to study the avian tree with high taxon sampling. The ASTRAL tree reported by [189] has low MLBS (Fig. 3.22), and many branches remain poorly supported with posterior probabilities (the median difference between BS and collapsed posterior was 0). Moreover, many of the most interesting relationships are poorly supported. For example, the sister to parrots+passerine has only 30% MLBS support, 83% local posterior support, and 0% collapsed posterior. These low support values are encouraging because the sister to parrots+passerine is likely recovered incorrectly in this tree, as most recent studies put falcons as sister to this group [105, 240, 111, 152]. Overall, despite its large taxon sampling, this dataset provides little resolution for the early Neoaves radiation using ASTRAL because of the insufficient gene count for this high level of ILS. Just like the avian genomic data, here I obtain a wall of short branches around the assumed rapid radiation of Neoaves (Fig. 3.22).

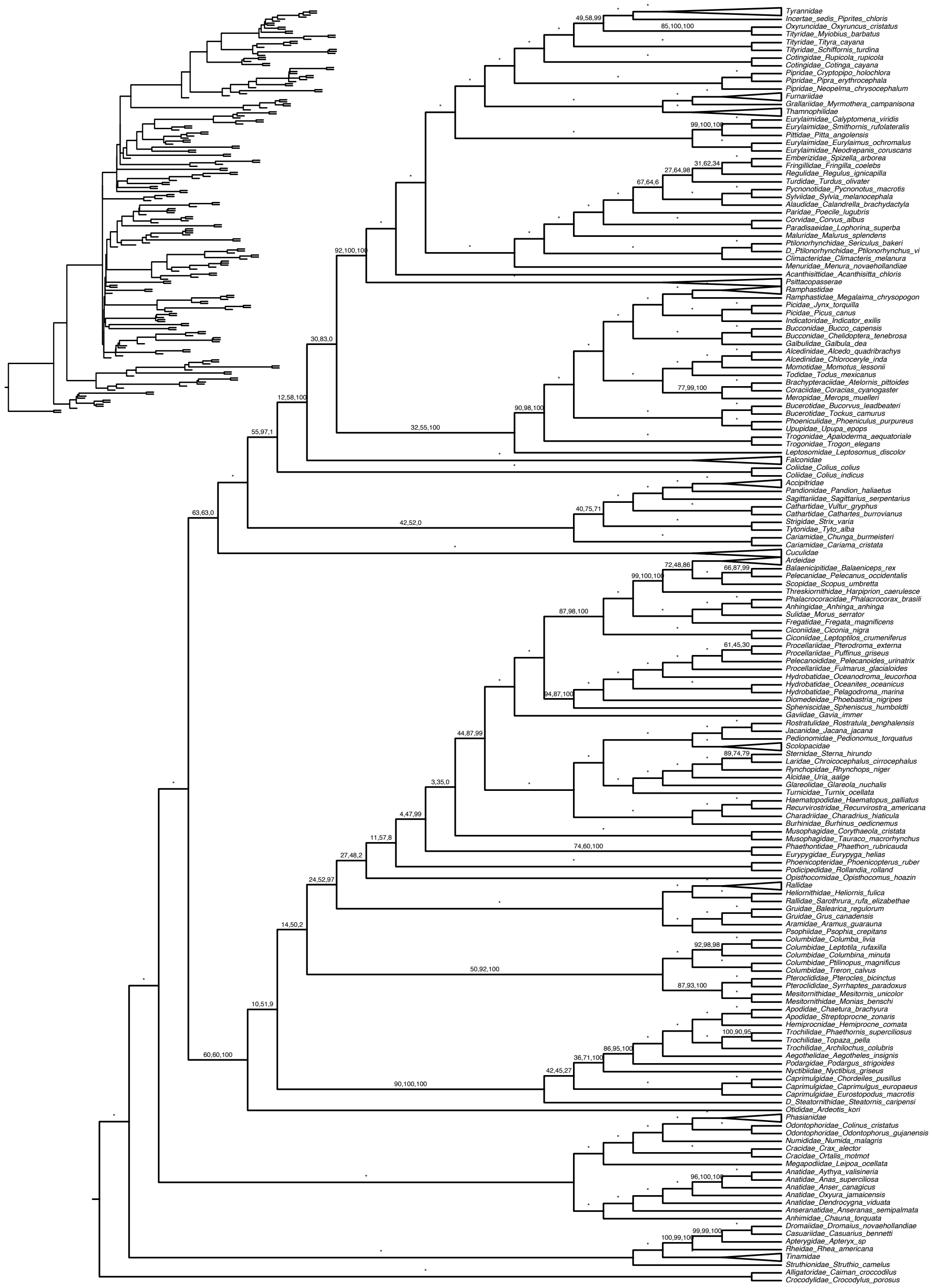

Tyrannidae
Incertae_sedis_Piprites_chloris
Oxyruncidae_Oxyruncus_cristatus
Tityridae_Myiobius_barbatus
Tityridae_Tityra_cayana
Tityridae_Schiffornis_turdina
Cotingidae_Rupicola_rupicola
Cotingidae_Cotinga_cayana
Pipridae_Cryptopipo_holochlora
Pipridae_Pipra_erythrocephala
Pipridae_Neopelma_chrysocephalum
Furnariidae
Grallariidae_Myrmothera_campanisona
Thamnophilidae
Eurylaimidae_Calyptomena_viridis
Eurylaimidae_Smithornis_rufolateralis
Pittidae_Pitta_angolensis
Eurylaimidae_Eurylaimus_ochromalus
Eurylaimidae_Neodrepanis_coruscans
Emberizidae_Spizella_arborea
Fringillidae_Fringilla_coelebs
Regulidae_Regulus_ignicapilla
Turdidae_Turdus_olivater
Pycnonotidae_Pycnonotus_macrotis
Sylviidae_Sylvia_melanocephala
Alaudidae_Calandrella_brachydactyla
Paridae_Poecile_lugubris
Corvidae_Corvus_albus
Paradisaeidae_Lophorina_superba
Maluridae_Malurus_splendens
Ptilonorhynchidae_Sericulus_bakeri
D_Ptilonorhynchidae_Ptilonorhynchus_vi
Climacteridae_Climacteris_melanura
Menuridae_Menura_novaehollandiae
Acanthisittidae_Acanthisitta_chloris
Psittacopasserae
Ramphastidae
Ramphastidae_Megalaima_chrysopogon
Picidae_Jynx_torquilla
Picidae_Picus_canus
Indicatoridae_Indicator_exilis
Bucconidae_Bucco_capensis
Bucconidae_Chelidoptera_tenebrosa
Galbulidae_Galbula_dea
Alcedinidae_Alcedo_quadribrachys
Alcedinidae_Chloroceryle_inda
Momotidae_Momotus_lessonii
Todidae_Todus_mexicanus
Brachypteraciidae_Atelornis_pittoides
Coraciidae_Coracias_cyanogaster
Meropidae_Merops_muelleri
Bucerotidae_Bucorvus_leadbeateri
Bucerotidae_Tockus_camurus
Phoeniculidae_Phoeniculus_purpureus
Upupidae_Upupa_epops
Trogonidae_Apaloderma_aequatoriale
Trogonidae_Trogon_elegans
Leptosomidae_Leptosomus_discolor
Falconidae
Coliidae_Colius_colius
Coliidae_Colius_indicus
Accipitridae
Pandionidae_Pandion_haliaetus
Sagittariidae_Sagittarius_serpentarius
Cathartidae_Vultur_gryphus
Cathartidae_Cathartes_burrovianus
Strigidae_Strix_varia
Tytonidae_Tyto_alba
Cariamidae_Chunga_burmeisteri
Cariamidae_Cariama_cristata
Cuculidae
Ardeidae
Balaenicipitidae_Balaeniceps_rex
Pelecanidae_Pelecanus_occidentalis
Scopidae_Scopus_umbretta
Threskiornithidae_Harpiprion_caerulesce
Phalacrocoracidae_Phalacrocorax_brasili
Anhingidae_Anhinga_anhinga
Sulidae_Morus_serrator
Fregatidae_Fregata_magnificens
Ciconiidae_Ciconia_nigra
Ciconiidae_Leptoptilos_crumeniferus
Procellariidae_Pterodroma_externa
Procellariidae_Puffinus_griseus
Pelecanoididae_Pelecanoides_urinatrix
Procellariidae_Fulmarus_glacialoides
Hydrobatidae_Oceanodroma_leucorhoa
Hydrobatidae_Oceanites_oceanicus
Hydrobatidae_Pelagodroma_marina
Diomedeidae_Phoebastria_nigripes
Spheniscidae_Spheniscus_humboldti
Gaviidae_Gavia_immer
Rostratulidae_Rostratula_benghalensis
Jacanidae_Jacana_jacana
Pedionomidae_Pedionomus_torquatus
Scolopacidae
Sternidae_Sterna_hirundo
Laridae_Chroicocephalus_cirrocephalus
Rynchopidae_Rhynchops_niger
Alcidae_Uria_aalge
Glareolidae_Glareola_nuchalis
Turnicidae_Turnix_ocellata
Haematopodidae_Haematopus_palliatus
Recurvirostridae_Recurvirostra_americana
Charadriidae_Charadrius_hiaticula
Burhinidae_Burhinus_oedicnemus
Musophagidae_Corythaeola_cristata
Musophagidae_Tauraco_macrorhynchus
Phaethontidae_Phaethon_rubricauda
Eurypygidae_Eurypyga_helias
Phoenicopteridae_Phoenicopterus_ruber
Podicipedidae_Rollandia_rolland
Opisthocomidae_Opisthocomus_hoazin
Rallidae
Heliornithidae_Heliornis_fulica
Rallidae_Sarothrura_rufa_elizabethae
Gruidae_Balearica_regulorum
Gruidae_Grus_canadensis
Aramidae_Aramus_guarauna
Psophiidae_Psophia_crepitans
Columbidae_Columba_livia
Columbidae_Leptotila_rufaxilla
Columbidae_Columbina_minuta
Columbidae_Ptilinopus_magnificus
Columbidae_Treron_calvus
Pteroclididae_Pterocles_bicinctus
Pteroclididae_Syrrhaptes_paradoxus
Mesitornithidae_Mesitornis_unicolor
Mesitornithidae_Monias_benschi
Apodidae_Chaetura_brachyura
Apodidae_Streptoprocne_zonaris
Hemiprocnidae_Hemiprocne_comata
Trochilidae_Phaethornis_superciliosus
Trochilidae_Topaza_pella
Trochilidae_Archilochus_colubris
Aegothelidae_Aegotheles_insignis
Podargidae_Podargus_strigoides
Nyctibiidae_Nyctibius_griseus
Caprimulgidae_Chordeiles_pusillus
Caprimulgidae_Caprimulgus_europaeus
Caprimulgidae_Eurostopodus_macrotis
D_Steatornithidae_Steatornis_caripensi
Otididae_Ardeotis_kori
Phasianidae
Odontophoridae_Colinus_cristatus
Odontophoridae_Odontophorus_gujanensis
Numididae_Numida_malagris
Cracidae_Crax_alector
Cracidae_Ortalis_motmot
Megapodiidae_Leipoa_ocellata
Anatidae_Aythya_valisineria
Anatidae_Anas_superciliosa
Anatidae_Anser_canagicus
Anatidae_Oxyura_jamaicensis
Anatidae_Dendrocygna_viduata
Anseranatidae_Anseranas_semipalmata
Anhimidae_Chauna_torquata
Dromaiidae_Dromaius_novaehollandiae
Casuariidae_Casuarius_bennetti
Apterygidae_Apteryx_sp
Rheidae_Rhea_americana
Tinamidae
Struthionidae_Struthio_camelus
Alligatoridae_Caiman_croccodilus
Crocodylidae_Crocodylus_porosus
49,58,99
85,100,100
99,100,100
31,62,34
27,64,98
67,64,6
92,100,100
30,83,0
77,99,100
90,98,100
32,55,100
12,58,100
55,97,1
40,75,71
42,52,0
63,63,0
72,48,86
66,87,99
99,100,100
87,98,100
61,45,30
94,87,100
44,87,99
89,74,79
3,35,0
4,47,99
11,57,8
74,60,100
27,48,2
24,52,97
14,50,2
92,98,98
50,92,100
87,93,100
10,51,9
100,90,95
86,95,100
36,71,100
42,45,27
90,100,100
60,60,100
96,100,100
99,99,100
100,99,100

Figure 3.22: Support on the avian dataset of Prum et al. [189] Three forms of support values are shown. Inset: branch lengths in coalescent units (terminal lengths are drawn arbitrarily). Collapsed clades had full support with all three methods for all branches.

3.5 Discussions

The local posterior probabilities introduced in this dissertation can be computed quickly and without a need for extensive MCMC sampling or bootstrapping. Computing posteriors for a species tree with 200 taxa and 1000 genes takes only 10 seconds and for a dataset with 1000 taxa and 1000 genes, takes about three minutes on a laptop machine. This extremely fast computation is possible only because of the two main assumptions of the method: that true MSC-generated gene trees are given and the locality assumption (i.e., the four clusters around each internal branch are present in the true tree). These assumptions can both be violated on real data. Recognizing this fact, my simulations include conditions that violate these two assumptions by introducing plenty of gene tree estimation error (ranging from average RF distance of 25% to 67%) as well as species tree error (Table 3.1 and Fig. 3.5).

My method allows very few false positives with high support, a pattern that is retained even with high levels of gene tree estimation error. It could be argued that my method is perhaps too conservative and underestimates support. Nevertheless, local posterior probabilities were *less* conservative than MLBS, the only viable alternative for large datasets. Despite allowing very few false positives with high support, the method generally had high recall (i.e., true branches with high support) except for very few genes for a given amount of ILS. Reassuringly, increased gene tree estimation error only negatively impacted recall but retained very high precision. While underestimation of support is not desirable, the abundance of false branches with high support would be a more serious problem.

A practical question is at what threshold of support a branch can be judged reliable. The answer depends on factors such as the false positive rate desired and the amount of gene tree error. Nevertheless, it seems that the commonly used 0.95 threshold results in very high precision while retaining moderately high recall. In my analyses, even lower thresholds (e.g., 0.9 or even 0.7) give high precision, while increasing the recall.

3.5.1 Gene tree estimation error

An interesting pattern was that with estimated gene trees (but not with true gene trees), at a given threshold, support values are more precise for high ILS compared to low ILS (Fig. 3.7A). I postulate this effect is related to the larger impact that gene tree estimation error has on the total amount of observed discordance for low ILS compared to high ILS conditions. Consistent with this explanation, I also observed a larger degradation of branch length accuracy in going from true to estimated gene trees for low ILS conditions compared to high ILS (Table 3.3).

The issue of gene tree estimation error is at the heart of why I saw a need for developing this new method. Sets of site-resampled bootstrap gene trees tend to have increased levels of discordance with regard to the species tree (and also among themselves) compared to ML gene trees, especially when each gene has a limited phylogenetic signal. Bootstrapped gene trees have much higher rates of discordance than either true gene trees or ML gene trees (Fig. 3.23). It is expected that bootstrapped replicates of a dataset result in noisier estimates of parameters than ML; however, the added error by bootstrapping should not be *biased*. For MLBS, the input to the summary method is not just a noisy dataset, but a *biased* one with increased levels of discordance. I postulate this bias is the reason for the underperformance of MLBS.

My method, on the other hand, does not require bootstrapping and uses the best available gene trees (e.g., estimated ML gene trees). While ML gene trees are still biased towards increased discordance (and hence reduced branch length), they are better than bootstrapped gene trees (Fig 3.23). The downside of my approach is that gene tree uncertainty is not considered directly. Thus, it was reassuring to see that my method remains precise with high gene tree estimation error. To account for gene tree estimation error, one can collapse the most poorly resolved branches in gene trees when computing support. As I show in

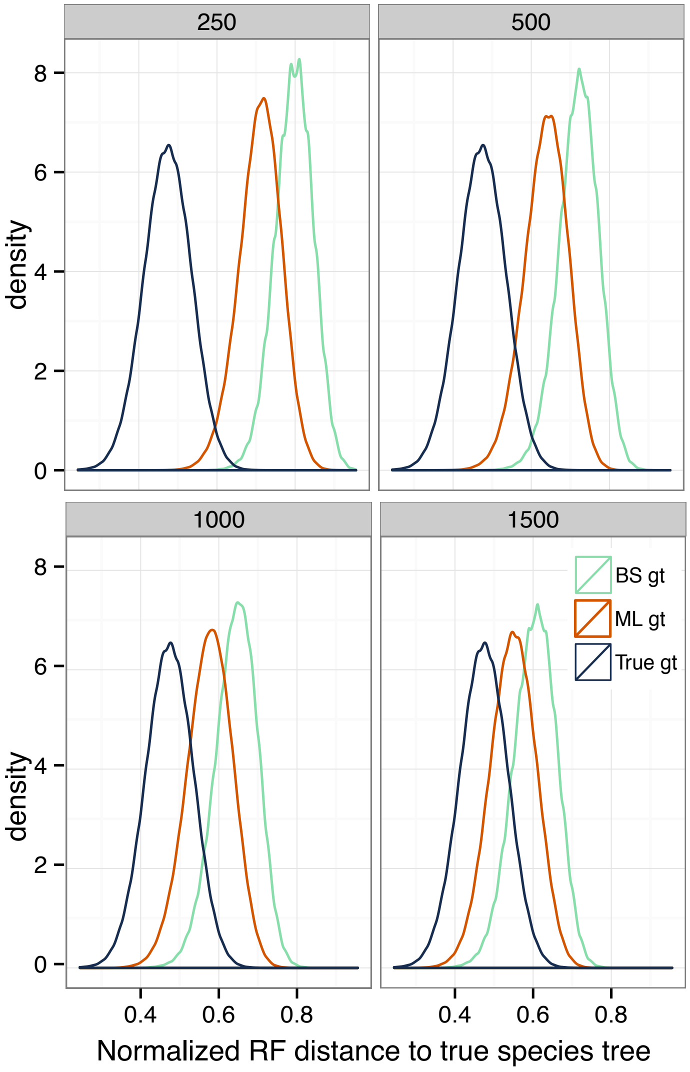


Figure 3.23: Gene tree discordance for the avian dataset. I show the density plot of the normalized RF distance between the true species tree and true gene trees, ML gene trees, and BS gene trees for fMy different model conditions.

my analyses of biological data, this practice seems promising. However, I note that when collapsing branches, one should be careful not to introduce bias, which can happen with aggressive filtering. I choose to collapse branches with support below 33% (which can be considered randomly resolved). Future work needs to further study the effect of collapsing low support branches in gene trees on both branch length and support.

The impact of gene tree estimation error was most clear with estimates of the branch length. The branch lengths produced by my method showed encouraging patterns (e.g., consistency across biological datasets); nevertheless, my estimated branch lengths are not immune to underestimation that is seen often with other summary methods. Thus, I suggest branch lengths from ASTRAL and other summary methods should be interpreted with care in the presence of gene tree estimation error.

3.5.2 High support despite high discordance

An important observation, predicted by the theory but sometimes lost in the scientific debate about discordance, is that high confidence for a correctly inferred relationship can emerge even with high levels of discordance. As Figure 3.3 shows, a branch that appears in only 40% of gene trees can still be resolved with high confidence if a sufficient number of genes are available (e.g., around 500). For example, in the 1KP tree, the branch that puts Zygnematales as sister to land plants appeared in only 49% of collapsed gene tree quartets and the branch making Bryophytes monophyletic only appeared in 50% of them; both branches, however, have a posterior of 1.0. I have implemented an option in ASTRAL to output the percentage of gene tree quartets that agree with each of the three resolutions around a branch. Pie charts in Figure 3.19 give examples of these relative quartet frequencies.

A question that biologists often face is the number of genes required to resolve a branch. The number of genes required to obtain high resolution and low false positive rates depends on the model condition. With higher ILS, more genes are required, an observation that is not surprising. However, my method can be extended to estimate the number of genes that might be required to resolve a tree (with an estimated level of ILS).

3.5.3 Limitations and future work

Promising approaches for incorporating gene tree uncertainty into local posterior probabilities exist. For example, one can weight each gene tree quartet around a branch by its SH-like support, BS, or advanced measures like concordance [11]. Moreover, comparing my method against Bayesian co-estimation methods on small datasets where they can run will be interesting. Furthermore, I did not investigate the impact of changing prior parameters (λ); nor did I explore other prior functions, such as Dirichlet distributions (conjugate to multinomial) or birth-death processes. I leave these for future work.

I violated some but not all assumptions of my method in the experimental results. The sequence evolution models used for simulation and inference were both GTR+Γ, but on real data, model violations (e.g., compositional bias) can lead to biased estimates of gene trees. Finally, all my simulated datasets had discordance that was generated only by ILS and estimation error, and not other sources of true biological discordance, such as undetected paralogy or horizontal gene transfer. Future work should further examine the impact of other biological sources of discordance on the reliability of local posterior probabilities.

3.6 Acknowledgements

Chapter 3, in full, contains material from SAYYARI, E., AND MIRARAB, S. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. Molecular Biology and Evolution 33, 7 (2016), 1654–1668. I was the primary investigator and author of this paper.

Chapter 4

Testing for polytomies in phylogenetic species trees using quartet frequencies

Phylogenetic species trees typically represent the speciation history as a bifurcating tree. Speciation events that simultaneously create more than two descendants, thereby creating *polytomies* in the phylogeny, are possible. Moreover, the inability to resolve relationships is often shown as a (soft) polytomy. Both types of polytomies have been traditionally studied in the context of gene tree reconstruction from sequence data. However, polytomies in the species tree cannot be detected or ruled out without considering gene tree discordance. in this dissertation, I describe a statistical test based on properties of the multi-species coalescent model to test the null hypothesis that a branch in an estimated species tree should be replaced by a polytomy. On both simulated and biological datasets, I show that the null hypothesis is rejected for all but the shortest branches, and in most cases, it is retained for true polytomies. The test, available as part of the Accurate Species TRee ALgorithm (ASTRAL) package, can help systematists decide whether their datasets are sufficient to resolve specific relationships of interest.

4.1 Introduction

Phylogenies are typically modeled as bifurcating trees. Even when the evolution is fully vertical, which it is not always [19, 171], the binary model precludes the possibility of several species evolving simultaneously from a progenitor species [95]. These events could be modeled in a multifurcating tree where some nodes, called polytomies, have more than two children. True polytomies have been suggested for several parts of the tree-of-life (e.g., [239, 14]). Polytomies are also used when the analyst is unsure about some relationships due to a lack of signal in the data to resolve relationships [251]. The terms *hard* and *soft* polytomies are used to distinguish between these two cases [146], with a *soft* polytomy reserved for the case where relationships are unresolved in an estimated tree and a *hard* polytomy for multifurcations in the true tree (Fig. 4.1). Distinguishing the two types of polytomies is not easy. Moreover, the distinction between soft and hard polytomies can be blurred. The difficulty in resolving relationships increases as branches become shorter. In the limit, a branch of length zero is equivalent to a hard polytomy, which is not just difficult but impossible to resolve. Regardless of abstract distinctions, a major difficulty faced by systematists is to detect whether specific resolutions in their inferred trees are sufficiently supported by data to rule out a polytomy (e.g., see [48, 239]).

For any branch of a given species tree, I can pose a *null* hypothesis that the length of that branch is zero, and thus, the branch should be removed to create a polytomy. Using observed data, I can try to reject this null hypothesis, and if I fail to reject, I can replace the branch with a polytomy. The resulting polytomy is best understood as a soft polytomy because the inability to reject a null hypothesis is never accepting the alternative hypothesis. The inability to reject may be caused by a real (i.e., hard) polytomy, but it may also simply be due to the lack of power (Fig. 4.1). in this dissertation, I present a new test with polytomy as the null hypothesis for multi-locus datasets.

The idea of testing a polytomy as the null hypothesis and rejecting it using data has been applied to single-locus data [103, 259]. Likelihood ratio tests against a zero-length branch have been developed (e.g., Swofford–Olsen–Waddell–Hillis (SOWH) test [246]) and are implemented in popular packages such as Phylogenetic Analysis Using Parsimony (PAUP*) [245]. Treating polytomies as the null hypothesis has been pioneered by Walsh *et al.* who sought to not only test for polytomies but also to use a power analysis to distinguish soft and hard polytomies [259]. Appraising their general framework, Braun and Kimball [34] showed that the power analysis can be sensitive to model complexity (or lack thereof). Perhaps most relevant to my work, Anisimova *et al.* presented an approximate but fast likelihood ratio test for a polytomy null hypothesis [12]; their test, like what I will present looks at each branch and its surrounding branches while ignoring the rest of the tree.

Existing tests that treat pose a polytomy as the null hypothesis assume that the sequence data are generated on a single tree, as do Bayesian methods of modeling polytomies [133]. Therefore, these methods test whether a *gene tree* includes a polytomy [222]. However, the species tree can be different from gene trees and the discordance can have several causes, including gene duplication and loss, lateral gene transfer, and incomplete lineage sorting [56, 147]. Arguably, the question of interest is whether the species tree includes polytomies. Moreover, we are often interested to know whether we should treat the relationship between species as unresolved given the amount of data at hand. These questions cannot be answered without considering gene tree discordance, an observation made previously by others as well [222, 187]. For example, Poe and Chubb [187], in analyzing an avian dataset with five genes, first looked for zero-length branches in the gene trees using the SOWH test [246] and found evidence that some gene trees may include polytomies. However, they also tested if the pairwise similarity between gene trees was greater than a set of random trees and it was not. Their test of gene tree congruence, however, was not with respect to any particular model of gene tree evolution.

A major cause of gene tree discordance is a population-level process called incomplete lineage sorting (ILS), which has been modeled by the multi-species coalescent (MSC) model [191, 178]. The model tells us that that likelihood of ILS causing discordance increases as branches become shorter; therefore, any test of polytomies should also consider ILS. The MSC model has been extensively used for reconstructing species trees using many approaches, including Bayesian co-estimation of gene trees and species trees [90, 139] and site-based approaches [47, 38]. A popular approach (due to its scalability) is the summary method, where I first reconstruct gene trees individually and then summarize them to build the species tree. Many approaches that model ILS rely on dividing the dataset into quartets of species. These quartet-based methods (e.g., the summary method ASTRAL [164, 165, 278], the site-based method SVDQuartets [47], and a hybrid method called Bucky-Quartet [125]) rely on the fact that for a quartet of species, where only three unrooted tree topologies are possible, the species tree topology has the highest probability of occurring in unrooted gene trees under the MSC model (Fig. 4.1a).

Relying on the known distribution of quartet frequencies under the MSC model, I previously introduced a way of computing the support of a branch using a measure called local posterior probability (localPP) [210]. in this dissertation, I further extend the approach used to compute localPP to develop a fast test for the null hypothesis that a branch has length zero. Under the null hypothesis, I expect that the three unrooted quartet topologies defined around the branch should have equal frequencies [7]. This can be rigorously tested, resulting in the approach I present. Similar ideas have been mentioned in passing previously by Slowinski [222] and by Allman *et al.* [7] but to my knowledge, these suggestions have never been implemented or tested. The statistical test that I present is implemented inside the ASTRAL package (option `-t 10`) since version 4.11.2 and is available online at `https://github.com/smirarab/ASTRAL`.

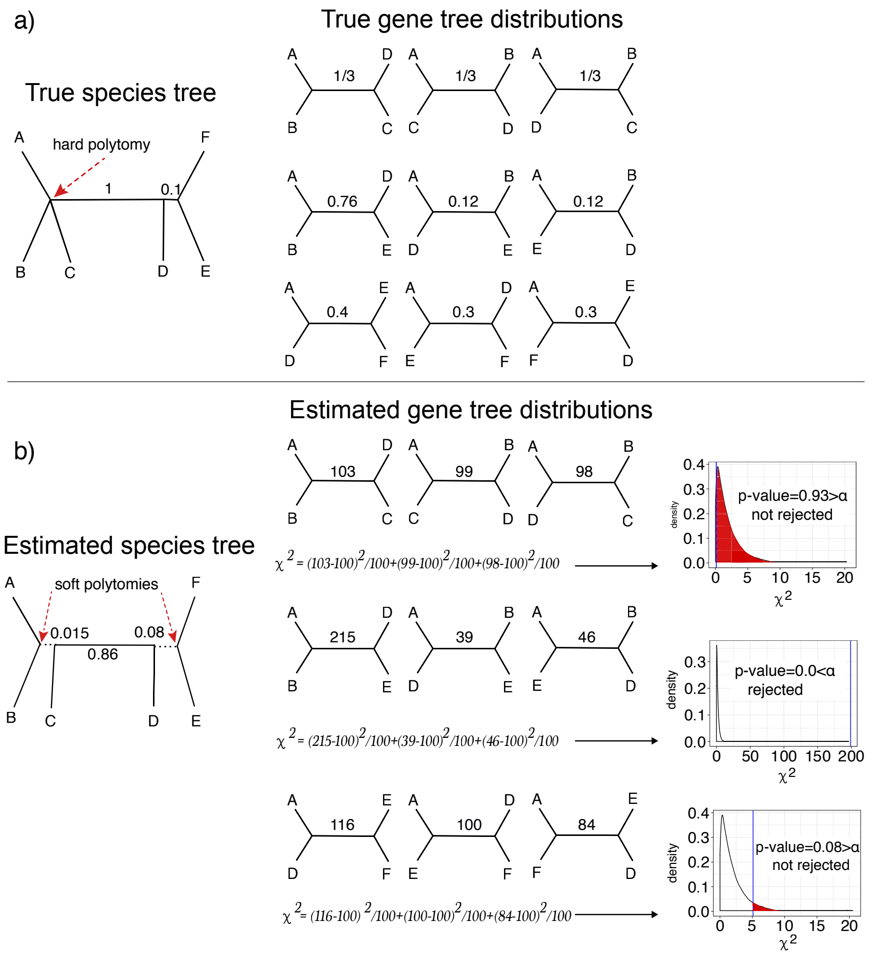


Figure 4.1: A statistical test of polytomies. (a) I show an example true species tree with a hard polytomy and two branches with CU lengths 1 and 0.1 (left), and the expected quartet gene tree fractions for three selected quartets based on the MSC model (right). The first quartet is around a true (hard) polytomy and has $\frac{1}{3}$ fraction expected for all three alternative topologies. (b) An estimated species tree with estimated CU branch lengths (left) and a hypothetical set of quartet tree frequencies counted from 300 hypothetical gene trees (right). Below each set of quartets, I show the computation of the χ^2 test statistic, and show where it falls on the χ^2 distribution with $DF = 2$; the vertical blue line shows the computed χ^2 based on given counts and the shaded red areas inside distributions show the area under the χ^2 distribution corresponding to the p-value. The null hypothesis is not rejected for the branch corresponding to the true polytomy (a true negative) or the short 0.1 CU branch (a false negative); these branches (dotted lines in the species tree) can be replaced with soft polytomies. CU: Coalescent unit; MSC: Multi-species coalescent. DF: Degrees of freedom

4.2 Materials and Methods

4.2.1 Background

An unrooted tree defined on a quartet of species $\{a,b,c,d\}$ can have one of three topologies (Fig. 4.1a): $t_1 = ab\|cd$ (i.e., a and b are closer to each other than c, and d), $t_2 = ac\|bd$, or $t_3 = ad\|bc$. Consider an unrooted species tree $ab\|cd$ where the internal branch length separating species a and b from c and d is x in coalescent units (CU), which is the number of generations divided by the haploid population size [56]. Under the Multi Species Coalescent (MSC) model, each gene tree matches the species tree with the probability $p_1 = 1 - \frac{2}{3}e^{-x}$ and matches each of the two alternative topologies with the probability $p_2 = p_3 = \frac{1}{3}e^{-x}$ [7]. Given true (i.e., error-free) gene trees with no recombination within a locus but free recombination across loci, frequencies n_1, n_2, n_3 of gene trees matching topologies t_1, t_2, t_3 will follow a multinomial distribution with parameters p_1, p_2, p_3 and with the number of trials equal $n = n_1 + n_2 + n_3$. Clearly, for a species tree with $N > 4$ species, the same results are applicable for any of the $\binom{N}{4}$ selections of a quartet of the species with x defined to be the branch length on the species tree restricted to the quartet (see Fig. 4.1 for examples).

4.2.2 A statistical test of polytomy

A true polytomy is mathematically identical to a bifurcating node that has at least one adjacent branch with length zero; the zero-length branch can be contracted in the binary tree to introduce the polytomy (e.g., compare branches P4 – P6 in Fig. 4.2a to the multifurcating tree), If the true species tree for a quartet of taxa has a polytomy (i.e., $x = 0$), then all gene tree topologies are equally likely with $p_1 = p_2 = p_3 = \frac{1}{3}$. Thus, if I had the true p_i values, I would immediately know if the species tree has a polytomy. However, I can never know true

p_i parameters; instead, I have observations n_1,n_2,n_3 with $\mathbb{E}(n_i)/n = p_i$. Luckily, multinomial distributions concentrate around their mean. As the number of genes increases, the probability of quartet frequencies deviating from their mean rapidly drops; for example, according to Hoeffding's inequality, the probability of divergence by ε drops exponentially and is no more than $2e^{-2\varepsilon^2 n}$. This concentration gives us hope that even though I never know true p_i values from limited data, I can design statistical tests for a $p = \frac{1}{3}$ null hypothesis. For an internal branch $\mathcal{B}$ in a bifurcating species tree, consider the following

Null hypothesis: The length of the internal branch $\mathcal{B}$ is zero; thus, the species tree has a polytomy.

To test this null hypothesis, I can use quartet gene tree frequencies given three assumptions.

A1. All positive length branches in the given species tree are correct.

A2. Gene trees are a random error-free sample from the distribution defined by the MSC model.

A3. I have $n \geq 10$ gene trees with at least a quartet relevant to $\mathcal{B}$.

A1, which I have previously called the locality assumption [210], can be somewhat relaxed. For each bipartition (i.e., branch) of the true species tree, either that bipartition or one of its NNI rearrangements should be present in the given species tree.

I now describe expectations under the null hypothesis. With start with the $N = 4$ case. By the A2 assumption, frequencies n_1,n_2,n_3 follow a multinomial distribution with parameters (p_1,p_2,p_3,n). Under the null hypothesis $p_1 = p_2 = p_3 = \frac{1}{3}$. Thus,under the null,

$$\chi^2 = \frac{(n_1 - n/3)^2}{n/3} + \frac{(n_2 - n/3)^2}{n/3} + \frac{(n_3 - n/3)^2}{n/3} \tag{4.1}$$

is asymptotically a chi-squared random variable with 2 degrees of freedom [276]. This chi-squared approximation for three equiprobable outcomes is a good approximation when $n \geq 10$ [115, 276, 193], hence my assumption A3. For smaller n's an exact calculation of the critical value is required [193], but I simply avoid applying my test for $n < 10$. Given the chi-squared random variable as the test statistic, I can simply use a Pearson's goodness-of-fit statistical test. Thus, the p-value is the area to the right of the χ^2 test statistic (Eq. 4.1) under the probability density function of the chi-square distribution with two degrees of freedom (Fig. 4.1b). This integral is available in various software packages, including the Java package `colt` [96], which I use.

With $N > 4$, I apply the test described above to each branch of the species tree independently. For each branch $\mathcal{B}$, I will have multiple quartets around that branch. I say that a quartet of species $\{a,b,c,d\}$ is *around* the branch $\mathcal{B}$ when it is chosen as follows: select an arbitrary leaf a from the subtree under the left child of $\mathcal{B}$, b from the subtree under the right child of $\mathcal{B}$, c from the subtree under the sister branch of $\mathcal{B}$, and d from the subtree under the sister branch of the parent of $\mathcal{B}$ (this can be easily adopted for a branch incident to root). Note that by assumption A1 (and its relaxed version), the length of the internal branch of the unrooted species tree induced down to a quartet around $\mathcal{B}$ is identical to the length of $\mathcal{B}$. Thus, under the null hypothesis, for any quartet around $\mathcal{B}$, I expect that the length of the quartet branch should be zero. Thus, any arbitrary selection of a quartet around the branch would enable us to use the same exact test I described before for $N = 4$.

Computing and summarizing quartet frequencies

Following the approach I previously used for defining localPP [210], I can also use *all* quartets around the branch. More precisely, let $n_{i,j}$ for $1 \leq i \leq 3, 1 \leq j \leq n$ be the number of quartets around the branch $\mathcal{B}$ that in gene tree j have the topology t_i and let $f_{i,j} = \frac{n_{i,j}}{n_{1,j}+n_{2,j}+n_{3,j}}$.

Then, I define

$$n_i = \sum_{j=1}^{n} f_{i,j} = \sum_{j=1}^{n} \frac{n_{i,j}}{n_{1,j} + n_{2,j} + n_{3,j}} . \tag{4.2}$$

Given these n_i values, I use the χ^2 test statistic as defined by Equation 4.1, just as before.

While I use Equation 4.2 mostly for computational expediency, my approach can be justified. Let $x_{i,j,k}, 1 \leq i \leq 3, 1 \leq j \leq n, 1 \leq k \leq m$ be an indicator variable that is 1 if and only if the quartet k around the branch $\mathcal{B}$ has the topology t_i in gene tree j. Let $m_{i,k} = \sum_j x_{i,j,k}$ be the number of gene trees where a quartet k has the topology t_i. Note that any quartet k around $\mathcal{B}$ can be chosen; thus, my hypothesis testing approach would work if I define $n_i = m_{i,k}$ for *any* k and use those n_i values in Equation 4.1. In particular, the quartet with the median $m_{i,k}$ is a valid and reasonable choice. Moreover, note that if all gene trees are complete, Equation 4.2 simplifies to $n_i = \text{mean}_k\, m_{i,k}$. I further assume that in the (unknown) distribution of $m_{i,k}$ values, the mean approximates the median. Thus, I approximate $n_i = \text{mean}_k(m_{i,k}) \approx \text{median}_k(m_{i,k})$. I use this approximation because, as it turns out, computing the mean is more computationally efficient than using the median.

It may initially seem that computing the n_i values requires computing $f_{i,j}$ values, which would require $O(N^4 n)$ running time. This would be too slow for large datasets. Computing the median quartet score also requires $O(N^4 n)$. However, the mean quartet score can be computed efficiently in $O(N^2 n)$ using the same algorithm that I have previously described for the localPP [210]. I avoid repeating the algorithm here but note that it is based on a postorder traversal of each gene tree and computing the number of quartets shared between the four sides of the branch $\mathcal{B}$ and each tripartition defined by each node of each gene tree. This traversal is adopted from ASTRAL-II [165].

When gene trees have missing data, the definition of $f_{i,j}$ naturally discards missing quartets. Similarly, if the gene tree j includes a polytomy for a quartet, it is counted towards neither of the three $n_{i,j}$ values and so, is discarded. Then, Equation 4.2 effectively assigns a quartet k missing/unresolved in a gene tree j to each quartet topology i proportionally to the number of present and resolved quartets in the gene tree j with the topology i; in other words, a missing $x_{i,j,k}$ is imputed to $\sum_k x_{i,j,k}/m_j$ where m_j is the number of quartets present and resolved in the gene tree j. A final difficulty arises when *none* of the quartets defined around $\mathcal{B}$ are present or if all of the present ones are unresolved in a multifurcating input gene tree. When this happens, I discard the gene tree for branch $\mathcal{B}$, reducing the number of genes n. Thus, the *effective* number of genes (i.e., effective n) can change from one branch to another based on patterns of gene tree taxon occupancy and resolution. Note that the A3 assumption is with respect to this effective number of gene trees and not the total number.

4.2.3 Evaluations

I examine the behavior of my proposed test, implemented in ASTRAL 5.5.9, on several simulated and empirical datasets on conditions that potentially violate assumptions A1 and A2.

The empirical datasets are a transcriptomic insect dataset [212], a genomic avian dataset [105] with “super” gene trees resulting from statistical binning [161], two multi-loci *Xenoturbella* datasets by Rouse *et al.* [203] and Cannon *et al.* [43], and a transcriptomic plant dataset [264] (Table 4.1). Since in the empirical data, the true branch length or whether a node should be a polytomy is not known, I will report the relationship between the estimated branch lengths and p-values. I will also randomly subsample gene trees to test how the amount of data impacts the ability to reject the null; for this, I focus on selected branches that have been difficult to resolve in the literature.

Simulated Datasets with Polytomies (S12A and S12B)

I simulated two datasets starting from two fixed species trees with 12 species (S12A: Fig 4.2a, S12B: Fig 4.2b). For both species trees, the tree height is 1.6M generations and

Table 4.1: Datasets. ref: the first publication to produce the dataset, max height (known only for simulated dataset): the height of the tree in number of generations (population size fixed to 2×10^5), *N*: number of species, *n*: the maximum number of genes, *R*: number of replicates, *qs*: average ASTRAL quartet score as a measure of gene tree discordance; computed using the true species tree and true gene trees for simulated and the estimated species tree and estimated gene trees for the biological datasets, *GE*: average distance between true and estimated gene trees (known for simulated dataset).

Type	Dataset	ref	max height	*N*	*n*	*R*	*qs*	*GE*
Biological	Aves	[105]		48	2022	1	0.64	
	insect	[212]		144	1478	1	0.72	
	plant	[264]		103	844	1	0.89	
	Xenoturbella	[203]		26	393	1	0.50	
	Xenoturbella	[43]		78	212	1	0.55	
Simulated	S12A	new	1.6M	12	1000	50	0.82	36%
	S12B	new	1.6M	12	1000	50	0.68	35%
	S201	[165]	10M	201	1000	100	0.94	25%
	S201	[165]	2M	201	1000	100	0.72	31%
	S201	[165]	500K	201	1000	97	0.49	47%

the population size is 2×10^5; thus, the tree height is 8 CU. The S12A species tree has two polytomies, each with three children, in addition to a short branch (P0) of length 0.2 coalescent units. The S12B tree has a polytomy with five children. For both S12A and S12b, Simphy [149] is used to simulate 50 replicates, each with 1000 gene trees. After generating the true gene trees, I used Indelible [72] and the GTR+Γ model of sequence evolution to simulate 250bp sequences down the gene trees. The GTR+Γ parameters are drawn randomly from Dirichlet distributions used in the ASTRAL-II paper (parameters are estimated from a collection of biological datasets [165]). I then used FastTree2 [188] to estimate gene trees from the sequence data. Both datasets have around 35% gene tree error, measured as the average RF distance between true and estimated gene trees (Table 4.1).

On this datasets, I score an arbitrary resolution of the true multifurcating species trees. Therefore, I can have both false positive errors (incorrectly rejecting the null for a polytomy) and false negative errors (failing to reject the null for a positive-length branch). I vary the number of genes between 20 and 1000 by randomly subsampling them and examine the distribution of *p*-values across all 50 replicates for each interesting branch using both true and estimated gene trees.

Simulated Datasets without Polytomies (S201)

I use a 201-taxon simulated dataset previously generated [165, 210]. Species trees are generated using the Yule process with a maximum tree height of 500K, 2M, or 10M generations and speciation rates of 10^{-6} (50 replicates per model condition) and 10^{-7} (another 50 replicates). The population size is fixed to 2×10^5 in all datasets. Thus, I have three conditions, each with 100 replicates and each tree includes 198 branches (59,400 branches in total). Branch lengths have a wide range (as I will see). The estimated gene trees on this dataset have relatively high levels of gene tree error (Table 4.1). Each replicates has 1000 gene trees, which I also randomly subsample to 50 and 200.

In this dataset, the true species trees are fully binary and therefore, the null hypothesis is never correct. Any failure to reject the null hypothesis is a false negative error. The inability to reject the null hypothesis should never be taken as accepting the null hypothesis because it can. simply indicate that the available data is insufficient to distinguish a polytomy from a short branch. An ideal test should be able to reject the null for long branches. However, for very short branches, failing to reject the null would be the expected behavior. It is worth contemplating the meaning of super short branches. For a haploid population size of 10^5, a branch length of 10^{-4} CU corresponds to only ten generations. One can argue that such short branches, for most practical purposes, can be considered a polytomy. Thus, false negative errors among super short branches could perhaps be tolerated.

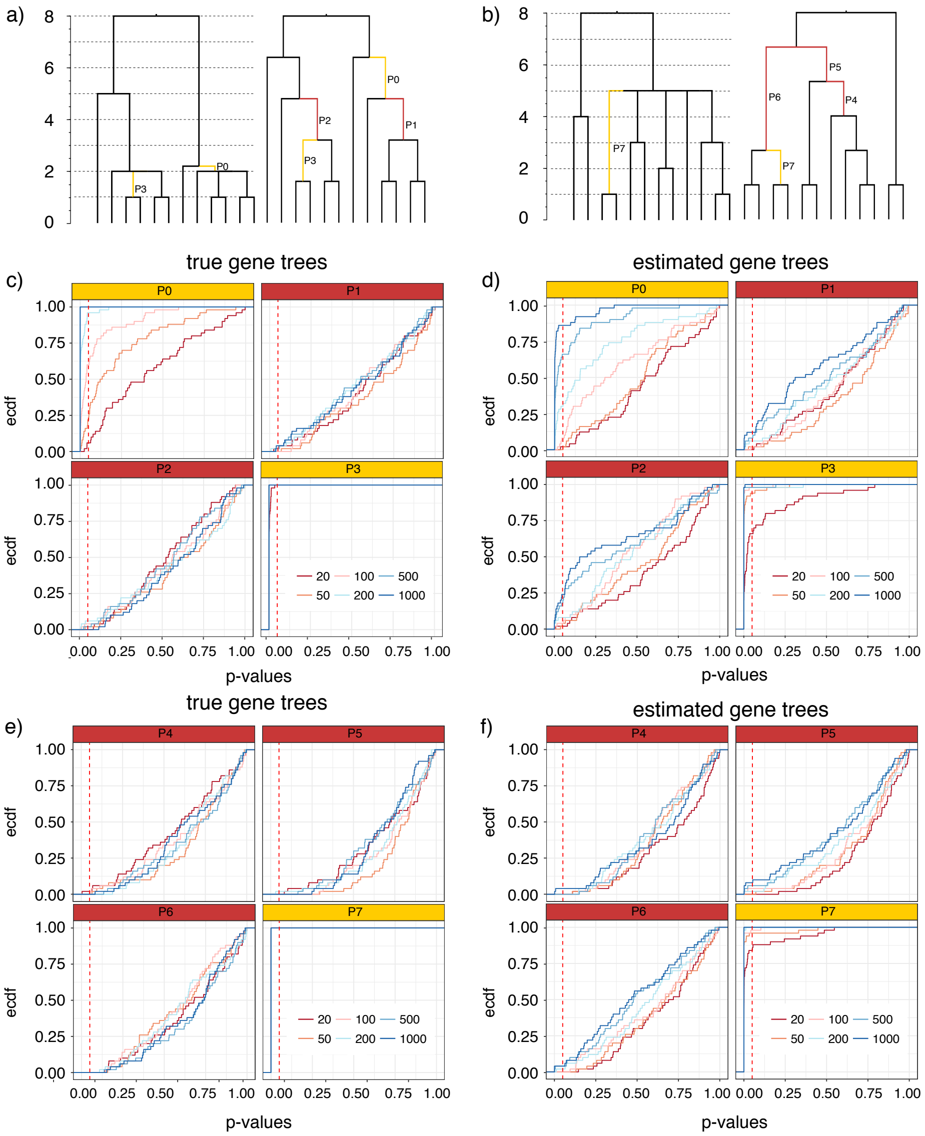


Figure 4.2: S12 datasets: true species trees and p-value distributions. For S12A (a) and S12B (b), I show the true multifurcating species trees in coalescent units (left) and an arbitrary resolution of the species tree used to test for polytomies (right). Branches P1, P2, and P4–P6 (red) represent arbitrary resolutions for which the null hypothesis is correct. Branches P0, P3, and P7 (yellow) are selected as examples for which the null hypothesis is incorrect. (c–f) The p-value distributions are shown as empirical cumulative distribution functions (ECDF) where the x-axis shows the p-value x and the y-axis shows the percentage of the replicates (out of 50) with a p-value $\leq x$. Results are shown for four selected branches of S12A (c,d) and S12B (e,f) for both true gene trees (c,e) and estimated gene trees (d,f) with varying numbers of gene trees (line colors). Dashed vertical red line shows p-value$= 0.05$. In red boxes, the intersection of the vertical line with each line shows the false positive rate. In yellow boxes, the intersection of the vertical line with each line shows one minus the false negative rate.

4.3 Results

4.3.1 Simulated datasets

I focus my discussions on $\alpha = 0.05$, but I show full distributions of p-values in many places.

S12A and S12B

On the S12A tree, P1 and P2 are zero-length branches and therefore, the test should ideally fail to reject the null hypothesis for them. As desired, when true gene trees are used, p-values are uniformly distributed (Fig. 4.2c; note the linear empirical cumulative distribution functions for P1 and P2 with true gene trees). For example, the null hypothesis is rejected for 4% of replicates with 1000 gene trees. As expected, since the null is correct, the false positive rate does not increase as I increase the number of gene trees. Switching to estimated gene trees universally increases false positive errors (Fig 4.2d). For example for P1, I reject the null hypothesis in 12% of replicates using 1000 gene trees. The most severe case of false positive error rates occurs for branch P2, where 24% of replicates are rejected with 1000 gene trees. Thus, gene tree errors can, in fact, increase the false positive error rates, but the extent of the increase depends on the length of branches surrounding the tested branch.

On the S12A tree, I also examine two binary positive-length branches: P0, which is short (0.2 CU length) and the parent of a polytomy, and P3 (1 CU), which is longer and the child of a polytomy. On these, I desire that the null hypothesis should get rejected. The P3 branch is easily rejected in all replicates using true gene trees. With estimated gene trees, given 50 genes or more, the null is rejected in almost all cases, and is rejected in 66% of replicates with 20 genes. Thus, the power to reject this moderate length branch (corresponding to 2×10^5 generations) is very high. For P0, which is rather short, the ability to reject the null hypothesis depends on the number of genes and similar to other branches, the power is higher for true gene trees. The false negative rates decrease as the number of genes increases; using 1000 gene trees, the null is rejected in all replicates with true gene trees and in 86% of replicates with estimated gene trees. Overall, the false negative rate is a function of the number of genes, the length of the branch, and gene tree error, as expected.

The S12B tree shows broadly similar results as S12A (Fig 4.2ef) but some differences are noteworthy. On the zero-length branches (P4, P5, and P6), as desired, the test fails to reject the null. However, false positives rates are a bit lower than expected by chance when true gene trees are used. For example, at $\alpha = 0.05$, I barely ever reject the null hypothesis for either of these three branches. These lower than expected false positive rates may be due to the fact that each branch is considered independently in my test, but P4, P5, and P6 are very much dependent (they all resolve one high degree polytomy). Even using estimated gene trees, the false positive rate remains low. With estimated gene trees, for P4, I reject the null in 4% of replicates when I use 1000 gene trees and I never reject the null hypothesis otherwise (Fig 4.2f). For P5 and P6, the false positive rates is at most 8% and 4% with 1000 genes. While the false positive rates remain low with estimated gene trees, the rate seems to slightly increase with increased numbers of gene trees. Alongside the zero-length branches, I also study the branch P7 (length: 2 CU), which is adjacent to the polytomy. For this relatively long branch, I always reject the null hypothesis with true gene trees. With estimated gene trees, the false negative rate is only 16% with 20 gene trees and gradually drops to 0% at 200 genes or more.

S201

On the S201 datasets, I can only have false negative errors. I bin branches according to the log of their CU length into 20 categories and compute the percentage of branches that are rejected according to my test with $\alpha = 0.05$ per bin (Fig 4.3). The false negative rate mostly depends on three factors: 1) the branch length, 2) the number of genes, and 3) whether

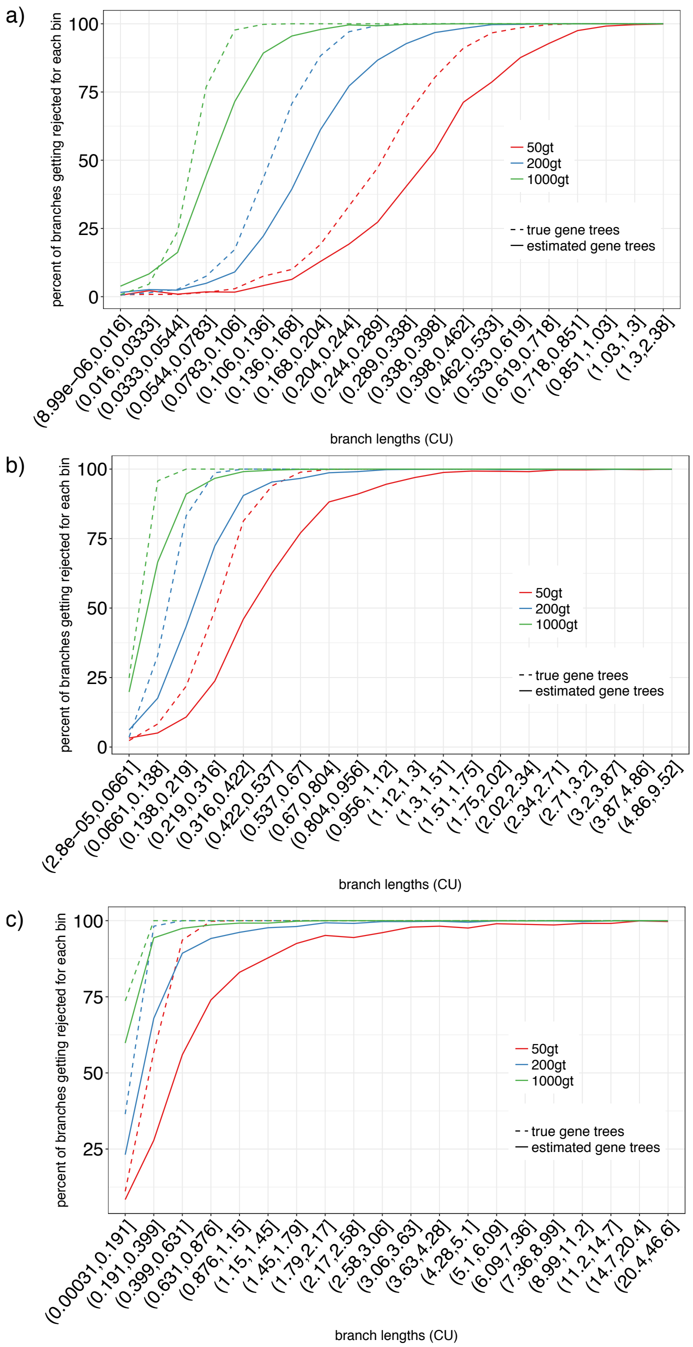


Figure 4.3: Polytomy test on S201 using estimated (solid) and true (dashed) gene trees for the different numbers of genes (colors) for model conditions with the tree height set to 500K (a), 2M (b), and 10M (c) generations. I show percentages of branches with the *p*-value ≤ 0.05 (y-axis) for branch length ranges (x-axis), formed by dividing the log of the true CU branch lengths into 20 equisized bins.

true or estimated gene trees are used. The impact of all three factors is consistent with what one would expect for a reasonable statistical test. For the longest branches (e.g., > 1.5 CU), the null hypothesis is rejected almost always even with as few as 50 genes and with my highly error-prone estimated gene trees. Using the true gene trees instead of estimated gene trees increases the power universally. For example, with 50 true gene trees, branches as short as 0.6 CU are almost always rejected. Interestingly, the difference between estimated and true gene trees seems to reduce as the number of genes increase.

Reassuringly, as the number of genes increases, the power to reject the null hypothesis also increases. Thus, with 1000 genes, branches between 0.1–0.2 CU are rejected 99.9% of the times with true gene trees and 90.0% of the times with estimated gene trees. Branches below $\log(7/6) \approx 0.15$ are considered very short and *can* produce the anomaly zone [55, 53]. Branches in the 0.05–0.15 CU range are rejected 90.4% and 67.4% of times with 1000 true and estimated gene trees, respectively.

4.3.2 Biological dataset

On the biological datasets, the ability to reject the null hypothesis depends on the branch length and the effective number of gene trees (Figure 4.4). Most branches with the estimated length greater than 0.1 CU had $p < 0.05$. Datasets with more than a thousand genes (Aves and insects) had higher resolution and have $p < 0.05$ for branches with the estimated length as lows as 0.035 CU. Yet in all datasets except the Aves (where all gene trees include all species) there are some ranges of branch length (often above 0.1 CU) where I am able to reject the null hypothesis for some branches but not for the others. This cannot be just due to random noise because estimated (not the unknown true) branch length is shown and two branches with the same length, have identical n_i/n values. Instead, the reason is that the effective number of genes changes from one branch to another because some gene trees may not include enough species to define a quartet around some branches. The effective number

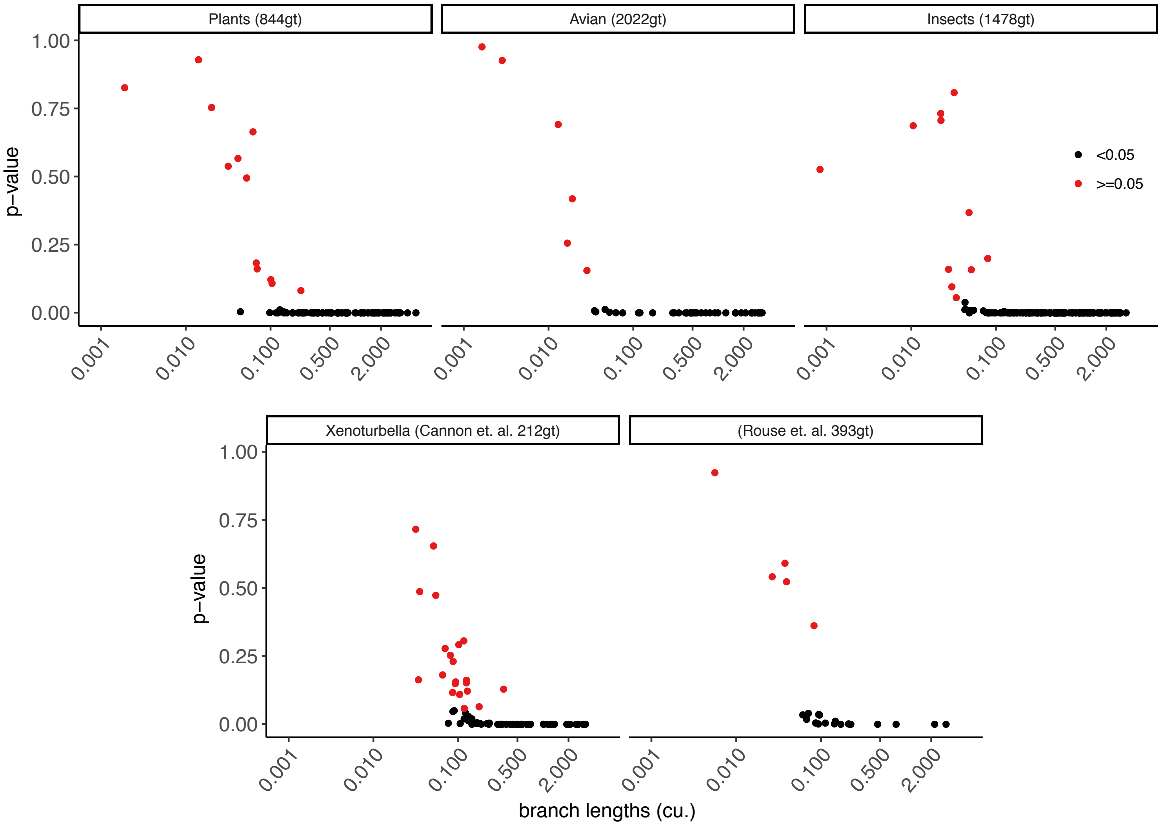


Figure 4.4: Polytomy test results for 5 different biological datasets using ASTRAL species trees, and all available gene trees. For each internal branch, I show its ASTRAL estimated CU length in log scale (x-axis) and its polytomy test *p*-value (y-axis). Points with $p < 0.05$ are in black. For each dataset (panel), the number of genes is reported inside the parentheses in the title.

of genes can also decline due to a lack of gene tree resolution, but this does not happen in my datasets, which include only binary gene trees (I will revisit this in the discussion section).

To further test the impact of the number of genes, for each dataset, I randomly subsampled gene trees (1% – 100% but no less than 20 gene trees) to find out how many genes are needed before I am able to reject the null hypothesis. I repeat this subsampling procedure 20 times, and show the average p-values across all 20 runs (Figures 4.5 and 4.6). In these analyses, I focus on selected branches of each empirical dataset. Note that in some downsampled datasets, occasionally branches have an effective number of genes that is smaller than 10, violating my assumption A3; I exclude these branches.

For the avian datasets, 6 branches in the species tree could not be rejected as a polytomy at $\alpha = 0.05$ even with all super gene trees (Figure 4.5a). These mostly belong to what has been called the wall-of-death [109], a hypothesized rapid radiation at the base of Neoaves [105, 180]. In subsampling super gene trees, I highlight seven selected branches (labeled A–G) as shown in Figure 4.5a. Interestingly, when I subsample super gene trees, several distinct patterns emerge for various branches. Most branches are easily rejected as a polytomy even with a small fraction of the data (e.g., C). For some shorter branches (e.g., G and B) rejecting a polytomy requires hundreds of super gene trees. Yet for others (e.g., D and perhaps F), I cannot reject the polytomy with the full dataset, but the pattern suggests that if I had more super gene trees, I may have been able to reject them as a polytomy. Finally, for some branches (e.g., A and G), increasing the number of genes does not lead to a substantial decrease in the p-value, suggesting that increasing the number of input trees may not be sufficient to resolve them.

For the insect dataset I focus on 6 clades, Holometabola, Acercaria+Hymenoptera, Hexapoda, Orthopteroidea, Pterygota, and Psocodea+Holometabola; these all have been classified as having *fairly strong support* from the literature [212], indicating that they enjoy robust support in the literature but some analyses reject them. As I reduce the number of

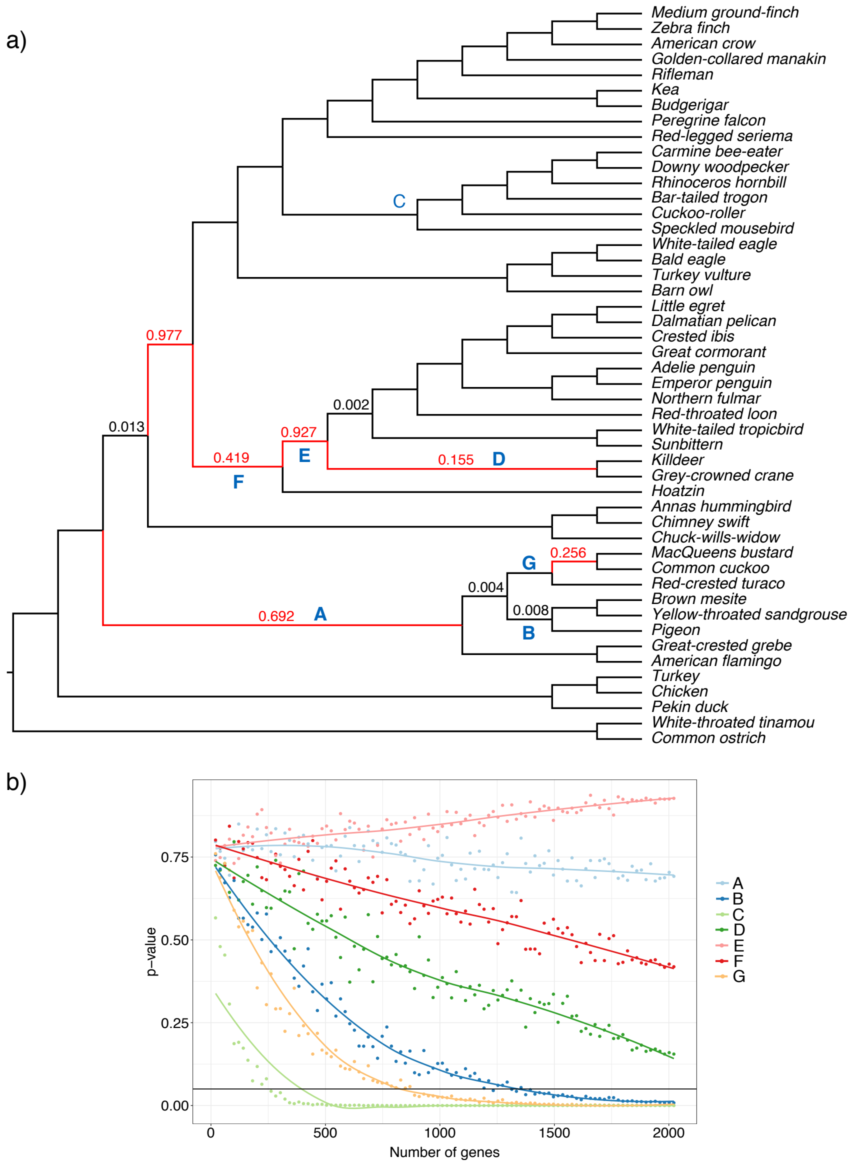


Figure 4.5: Polytomy test results on avian dataset. (a) ASTRAL species tree using binned ML gene trees. p-values greater than zero are reported on the branches, branches with $p > 0.05$ are shown in red. (b) change in p-value with respect to the number of genes for the selected branches in the species tree (labeled in blue in panel a). I used ASTRAL species trees with the varying number of gene trees sampled uniformly (1%, 2%, 3%,. . . , 100% of gene trees but no less than 20), and repeated 20 times. I show average p-values (y-axis) versus the number of gene trees (y-axis). Solid horizontal line shows p-value= 0.05.

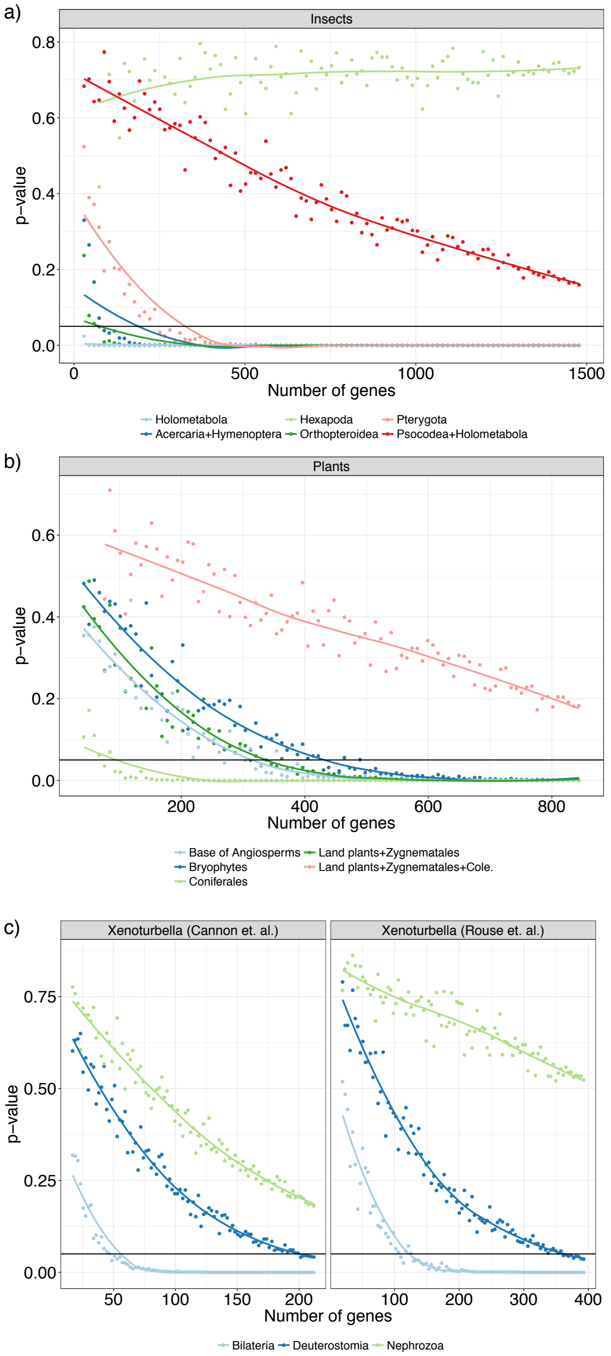


Figure 4.6: Polytomy test results for selected branches of (a) insects (b) plants, and (c) two *Xenoturbella* datasets. I used ASTRAL species trees with the varying number of gene trees sampled uniformly at random (1%, 2%,. . . , 100% of gene trees but no less than 20) repeated 20 times. I show average p-values (y-axis) versus the number of gene trees (x-axis). Solid horizontal line shows p-value$= 0.05$. Cases with effective n below 10 are excluded; for plants and *Xenoturbella*, I omit 1–4% because most replicates have $n < 10$.

genes, just like the avian dataset, I see three patterns (Figure 4.6a). For clades Holometabola, Acercaria+Hymenoptera, and Orthopteroidea, I get $p < 0.05$ even with fewer than 100 genes and for Pterygota with around 250 genes. I am not able to reject the null hypothesis for Psocodea+Holometabola with all gene trees, but the decreasing *p*-values suggest that this resolution could perhaps be resolved if I had several hundred more loci. The support for Hexapoda never decreases as I use more genes, suggesting that the relationship between insects and their close relatives (Collembola and Diplura, both considered insects in the past) may remain unresolved if I simply increase the number of genes. For this deep (around 450M years old) and undersampled node, *p*-values may fail to reduce either because of a true polytomy or because gene trees are estimated with high (perhaps biased) error.

In remaining datasets (plants and *Xenoturbella*), all important branches that I studied saw decreasing *p*-values as the number of gene trees increase (Figure 4.6b). In the plant dataset, having around 400 genes seems sufficient for most branches of interest, including the monophyly of Bryophytes and the resolution of Amborella as sister to all the remaining flowering plants. The branch that puts Zygnematales as sister to land plants is rejected as a polytomy with about 350 genes. However, the correct relationship between Chara and Coleochaetales remains hard to resolve. Even with the full dataset, a polytomy is not rejected, though the decreasing *p*-values point to the possibility that this relationship would have been resolved had I had more genes.

The *Xenoturbella* datasets both have three focal branches, surrounding the position of Xenacoelomorpha. The branch labeled Bilateria, which has Xenacoelomorpha and Nephrozoa as daughters branches in both papers, can be resolved at $\alpha = 0.05$ with as few as 50 (Cannon) or 100 (Rouse) gene trees (Figure 4.6c). However, pinpointing the position of Xenacoelomorpha also depends on the branch labeled Nephrozoa, which puts Xenacoelomorpha as sister to a clade containing Protostomia and Deuterostomia. The null hypothesis that this branch may be a polytomy is not rejected in either dataset, but a pattern of decreasing *p*-values with more loci can be discerned. Thus, both datasets are best understood as leaving the relationships between Protostomia, Xenacoelomorpha, and the rest of Deuterostomia as uncertain with some evidence that Xenacoelomorpha is at the base of Nephrozoa. Remarkably, patterns of difficulty in resolving branches are similar across the two independent datasets with different taxon and gene selection.

4.4 Discussion

I introduced a new test for rejecting the null hypothesis that a branch in the species tree should be replaced by a polytomy. Unlike existing tests, my new test considers gene tree discordance due to ILS, as modeled by the MSC model. In several simulations, I showed that the test behaves as expected. The null hypothesis is often retained for true polytomies and is often rejected for binary nodes, unless when the true branch lengths are very short. The power to reject the null hypothesis for binary relationships increases with longer branches or with more gene trees and is reduced with gene tree estimation error. Gene tree error can also, in some cases, increase the false positive rate.

4.4.1 Power

Overall, even when I have 1000 genes, it is rare that I can reject the null for branches shorter than 0.03 CU. A branch of length 0.03 corresponds to 6000 generations in my simulations. One can argue that failing to resolve a branch that corresponds to such short evolutionary times (roughly 60K years with a generation time of 10 years) can perhaps be tolerated. Mathematically, given a sufficiently large unbiased sample of gene trees, even infinitesimally short branches can be distinguished from a polytomy. In practice, however, extremely short branches should be treated with suspicion as my input gene trees invariably are not perfect samples from the MSC distribution.

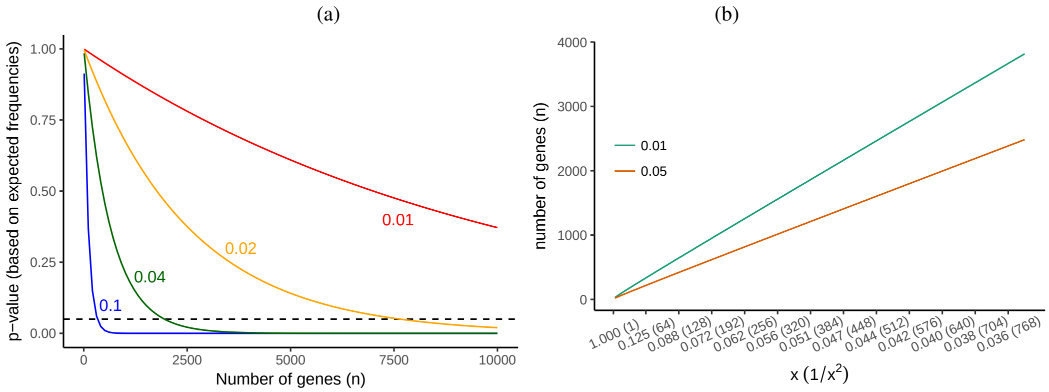


Figure 4.7: Impact of the number of genes on p-value (a) The p-value computed for the different number of gene trees (x-axis) for four different short branch lengths (colors) when the observed frequencies exactly match the expected frequencies given that branch length. Dashed horizontal line shows p-value$= 0.05$; it intersects at 331 for 0.1 CU, 1949 for 0.04 CU, 7641 for 0.02 CU, and 30259 for 0.01 CU (not shown). (b) The required number of genes (y-axis) to reject the null hypothesis with a p-value of 0.05 or 0.01 for various branch lengths (x-axis) assuming that the observed frequencies match the expected frequencies. Note that the x-axis scales with $\frac{1}{x^2}$.

In my biological analyses, I saw that subsampling genes and tracking trajectories of the p-value may be helpful in predicting the number of required genes to resolve a branch. The approach I presented can be used in other biological data as well. However, I caution that such predictions should be interpreted with the limitations of my proposed test in mind. Many factors such as gene tree error and other sources of discordance can contribute to deviations from MSC, and such deviations may render the predictions inaccurate. But if such predictions are to be made, a natural question arises: How does the number of genes impact the power?

I can easily compute the required number of genes for rejecting the null hypothesis assuming the expected frequencies match observed frequencies (Fig. 4.7a). For example, while for a branch of length 0.1 CU I only need $\approx$300 genes before I can reject it as a polytomy, for a branch of length 0.02 (i.e., 5 times shorter), I need $\approx$7500 genes (i.e., 25 times more). For a quartet species tree, $n_1 > \max(n_2, n_3)$ with arbitrarily high probability if the number of genes grows as $\frac{\log N}{x^2}$ [216]. More broadly, the number of genes required for correct species tree estimation using ASTRAL is proven to grow proportionally to $\log N$ and to x^{-2} [216].

Similarly, for any given branch length, I can numerically compute the minimum number of required genes to obtain a given p-value (e.g., 0.05). Assuming the observed frequencies match the expectations, I observe that the required number of genes grows linearly with $\frac{1}{x^2}$ (Fig. 4.7b). In fact, Figure 4.7b gives us a way to estimate the level of "resolution" that a dataset can provide. For example, 300 genes can reject the null for branches of $\approx$0.1 CU but if I quadruple the number of genes to 1200, my resolution is increased two-folds to branches of $\approx$0.05 CU. Note that gene tree error would increase these requirements and hence these should be treated as ballpark estimates. These estimates also assume I have $N = 4$ species.

The test I presented has no guarantees of maximal power. Other tests, such as likelihood ratio, may be more powerful. Moreover, it can be argued that my test is conservative in how it handles $N > 4$. When multiple quartets are available around a branch, I use their fraction supporting the $\mathcal{B}$ topology as the contribution of that gene to n_1. Thus, whether I have one quartet or a hundred quartets, I count each gene tree as one observation of my multinomial distribution. This is the most conservative approach to deal with the unknown dependencies between quartets. The most liberal approach would consider quartets to be fully independent, increasing the degrees of the freedom of the chi-square distribution to $2m$ instead of 2. Such a test would be more powerful but would be based on invalid independence assumptions that may raise false positive rates. An ideal test would need to model the intricate dependence structure of quartets, a task that is very difficult [65].

Finally, note that my test of polytomy relates to branch lengths in coalescent units. A branch of length zero in coalescent units will have length zero in the unit of time (or generations) if I keep the population size fixed. Mathematically, I can let the population size grow infinitely. For a mathematical model where the population size grows asymptotically faster than the time, one can have branches that converge to zero in length even though the branch length in time goes to infinity. This is just a mathematical construct with no biological meaning. Nevertheless, it helps to remind us that a very short branch in the coalescent unit (which my test may fail to reject as a polytomy) may be short not because the time was short but because the population size was large. Branches between 0.1 and 0.2 CU were not rejected as a polytomy by my test $\approx$10% of times even with 1000 genes. A length of 0.1 CU *can* correspond to 10M generations if the haploid population size is 100M.

4.4.2 Divergence from the MSC model and connections to localPP

The p-value from my proposed polytomy test has a close connection to the localPP branch support. Both measures assume the MSC model and both are a function of quartet scores (i.e., n_i/n). As the quartet score of the species tree topology and the number of genes increases, both localPP and $1-p$-value increase (Fig. 4.8a). When localPP of a branch is close to 1.0, the polytomy null hypothesis is always rejected. However, the two measures are not identical. Interestingly, there are some conditions where localPP is higher than 0.95 but the polytomy null hypothesis is not rejected at the 0.05 level (Fig. 4.8a). When the frequencies follow expectations of the MSC model, $1-p$-value of the polytomy test is smaller than the localPP.

It is important to remember that my test relies on the properties of the MSC model. If observed quartet frequencies diverge from the expectations of the MSC model systematically (as opposed to by natural variation), the behavior of my proposed test can change. For example, if n_2 is substantially larger than n_3, rejecting the null hypothesis becomes easier (Fig. 4.8b). This should not come as a surprise because this type of deviation from the MSC model makes the quartet frequencies even more diverged from $\frac{1}{3}$ than what is expected under the MSC model. On real data, several factors can may contribute to deviations from MSC. For example, incorrect homology detection in real datasets is possible (e.g., see [230] for possible homology issues with the avian dataset) and can lead to deviations.

Another source of deviation is gene flow, which can impact the gene tree distributions. Solís-Lemus *et al.* have identified anomaly zone conditions where the species tree topology

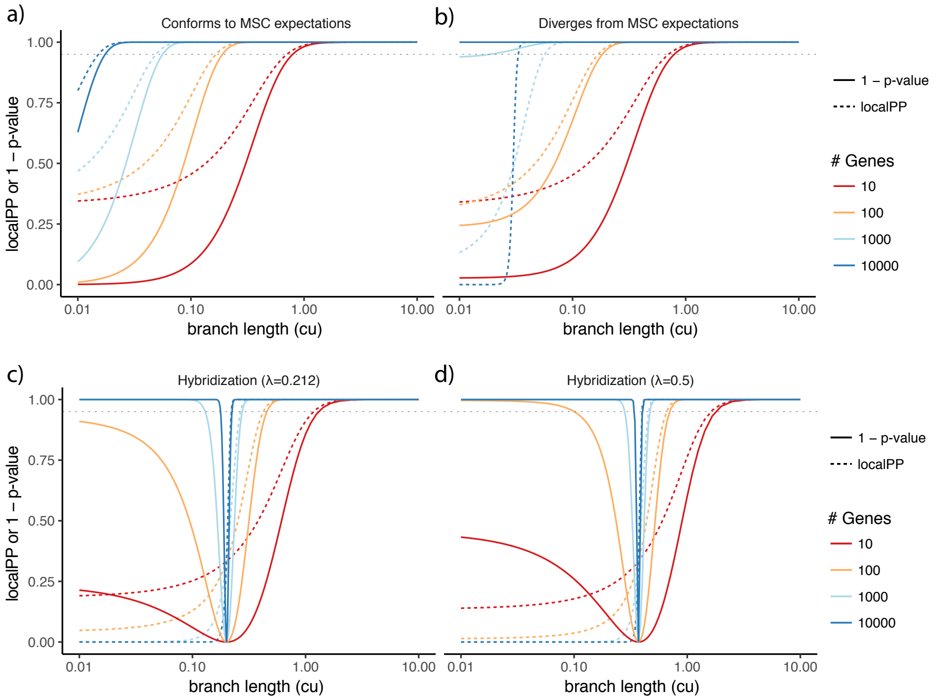


Figure 4.8: The polytomy test versus localPP. For various branch lengths (x-axis; log scale) and various numbers of gene trees (colors), I show (y-axis) both the localPP (dashed line) and $1-p$-value of the polytomy test (solid line). (a) The quartet frequencies follow MSC expectations: $\frac{n_1}{n} = 1-\frac{2}{3}e^{-x}$, $\frac{n_2}{n} = \frac{n_3}{n} = \frac{1}{2}\frac{2}{3}e^{-x}$. (b) The quartet frequencies diverge from the MSC expectations so that n_2 is 20% larger than n_3. $\frac{n_1}{n} = 1-\frac{2}{3}e^{-x}$, $\frac{n_2}{n} = \frac{6}{11}\frac{2}{3}e^{-x}$, $\frac{n_3}{n} = \frac{5}{11}\frac{2}{3}e^{-x}$. (c,d) quartet frequencies follow the MSC+gene flow model, as analyzed by Solís-Lemus *et al.* [225]. For a species tree with a hybridization at the base (see Fig. 2 of [225]) with inheritance probabilities $\lambda = 0.1$ (c) and $\lambda = 0.5$ (d), following Solís-Lemus *et al.*, I set $\frac{n_1}{n} = (1-\lambda)^2(1-\frac{2}{3}e^{-x}) + 2\lambda(1-\lambda)(1-e^{-x/2}+\frac{1}{3}e^{-x-4}) + \lambda^2(1-\frac{2}{3}e^{-x/2})$ and $n_2 = n_3 = \frac{n-n_1}{2}$. The dotted horizontal gray line shows p-value$= 0.05$.

has lower quartet frequencies compared to the alternative topologies [225]. Since the localPP measure does not model gene flow, under those conditions, it will be misled, giving low posterior probability to the species tree topology in the presence of gene flow (Fig. 4.8c). For example, if $\lambda = 0.1$ (meaning that 10% of genes are impacted by the horizontal gene flow), for branches of length 0.1 or shorter, localPP will be zero. The presence of the gene flow also impacts the test of the polytomy. For the species tree defined by Solís-Lemus *et al.* (Fig. 1 of [225]), when internal branches are short enough, there exist conditions where the gene flow and ILS combined result in quartet frequencies being equal to $\frac{1}{3}$ for all the three alternatives. It is clear that my test will not be able to distinguish such a scenario from a real polytomy (Fig. 4.8cd). One is tempted to argue that perhaps high levels of gene flow between sister branches *should* favor the outcome that the null is not rejected. However, this argument fails to explain the observation that for any value of λ, the null hypothesis is retained only with very specific settings of surrounding internal branch lengths (Fig. 4.8cd). Thus, I simply caution the reader about the interpretation when gene flow and other sources of bias are suspected.

4.4.3 The effective number of gene trees

It is important to note that the effective number of gene trees (effective-*n*) can change across branches of the same species tree. Missing data can reduce the number of genes that have at least one taxon from a quartet defined around the branch of interest. In my biological datasets, various branches of the same dataset often have a wide range of effective *n* (Fig 4.9a), especially for the two transcriptomic datasets (insects and plants) with lots of missing data. The only exception is the avian dataset, where my super gene trees always include all the taxa.

A second factor that can reduce the effective *n* is multifurcations in input gene trees. If all the quartets around a branch are unresolved in an input gene tree, that gene tree does not count towards the effective *n*. My biological datasets had binary gene trees. However, as recently shown [278], removing branches with very low support can help addressing gene

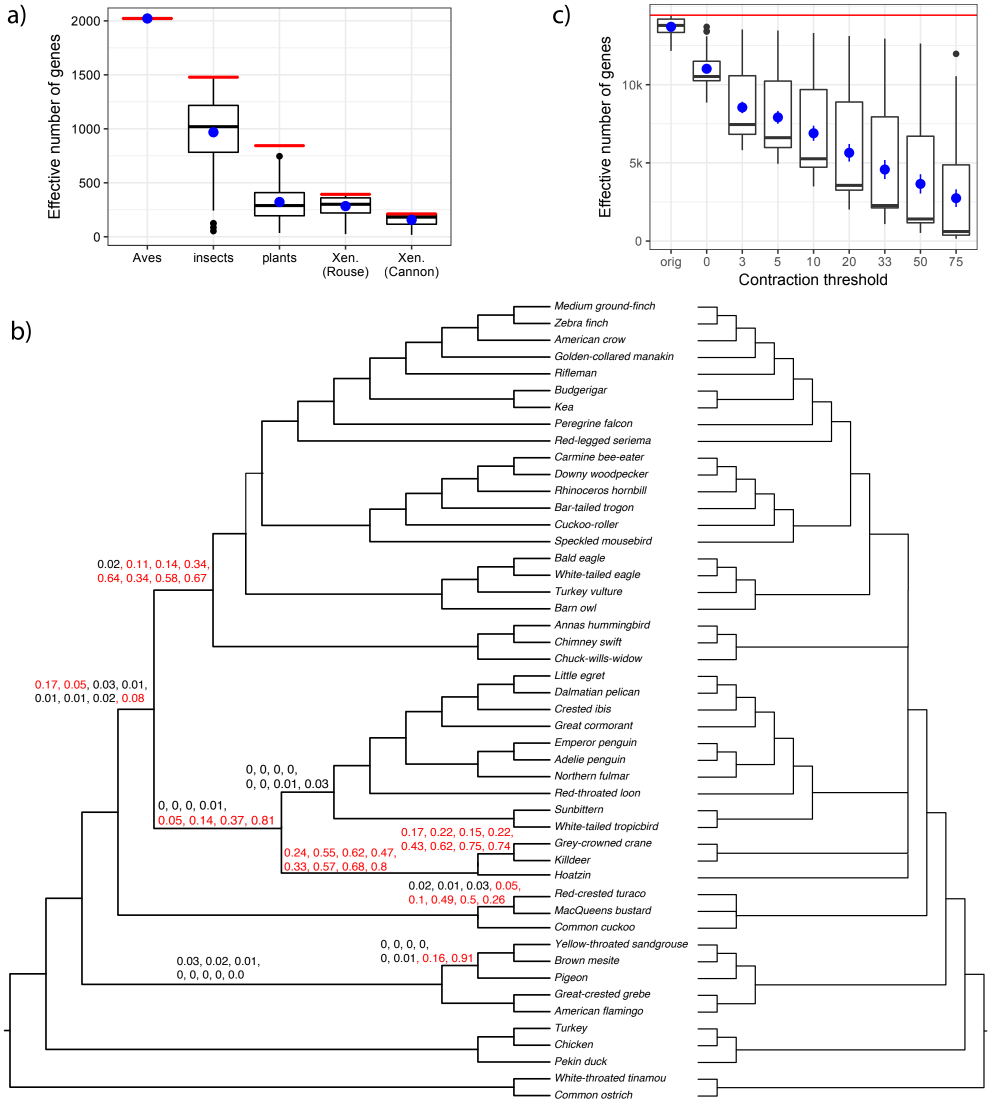


Figure 4.9: Effective *n* and results on the unbinned avian dataset. (a) Distributions of effective *n* (y-axis) across different branches of each empirical dataset (x-axis). I show boxplots (black) as well as mean and standard error (blue). The total number of genes (*n*) is shown as a red horizontal line. (b) ASTRAL-III species tree estimated based on 14,446 unbinned gene trees with branches up to 10% support contracted. For each branch, I show eight *p*-values that are computed, respectively, with respect to gene trees where branches with support up to 0%, 3%, 5%, 10% (top), 20%, 33%, 50%, or 75% (bottom) are contracted. Branches with no values have only 0 *p*-values (to three decimal points). *p*-values above 0.05 are in red. I also show the multifurcating species tree where all five branches that have *p*-values< 0.05 according to the 10% threshold are contracted (the left facing tree). (c) Similar to (a), I show distributions of effective *n* (y-axis) across branches of the avian species tree with all 14,446 original unbinned trees (orig) or with gene tree branches with low support contracted (x-axis).

tree error. To demonstrate this, I revisit the avian dataset. The purpose of using super gene trees instead of normal (unbinned) gene trees was to reduce the gene tree estimation error. Alternatively, one can simply remove branches with support at or below a certain threshold in gene trees and use the resulting tree as input to ASTRAL [278]. With this procedure and the support threshold set to 10%, I generated a new ASTRAL tree based on all 14,446 unbinned gene trees from the avian dataset [105, 161] (Fig 4.9b). The resulting tree was largely congruent with the ASTRAL tree on super gene trees and with reference phylogenies form the original publication [105].

I tested how the effective *n* and *p*-values change as a result of contracting low support branches. Simply contracting branches with 0% support reduces the median effective *n* from 13,791 to 10,523. Further contracting branches with support up to 3% – 75% gradually reduces the effective *n* all the way to a median of 610 (Fig 4.9c). The *p*-values tend to decrease as I increase the threshold for contraction (Fig 4.9b). Several branches fail to reject the null hypothesis regardless of the threshold chosen. Others reject the null hypothesis with lower levels of contraction but not with the higher levels, showing that the reduced effective *n* can reduce the power. For one branch, interestingly, the null is not rejected if I contract up to 0% and 3% support *or* if I contract up to 75%, but is rejected otherwise. This pattern may have a subtle explanation. With gene trees that include low support branches (up to 3%), I am unable to reject the null hypothesis perhaps because gene tree error creates a uniform distribution of quartets around this branch. As I further remove low support branches from the gene trees, I start to see quartet frequencies that favor the ASTRAL resolution perhaps because noise is removed and the actual signal can be discerned. Finally, with aggressive filtering of gene tree branches, effective *n* becomes so low that the test simply does not have the power to reject the null. These interesting patterns suggest that dealing with gene tree error by contracting low support branches may be possible, but the choice of the best threshold is not obvious. Future studies should further consider this question.

4.4.4 Interpretation

In the light of the dependence of my test on the MSC properties, I offer an alternative description of the test. A safe way to interpret the results of the test, regardless of the causes of gene tree discordance, is to formulate the null hypothesis as follows.

Null hypothesis: The estimated gene tree quartets around the branch $\mathcal{B}$ support all three NNI rearrangements around the branch in equal numbers.

This is the actual null hypothesis that I test. Under my assumptions, this hypothesis is equivalent to branch $\mathcal{B}$ being a polytomy. Under more complex models, such as gene flow + ILS, this null hypothesis holds true for polytomies but also for some binary networks.

The judicious application of my test will preselect the branches where a polytomy null hypothesis is tested and examines the p-value only for those branches. When many branches are tested, one arguably needs to correct for multiple hypothesis testing, further reducing the power of the test. Corrections such as Bonferroni or FDR [23] can be employed (but I did not apply them in my large scale tests that did not target specific hypotheses). However, note that even though I formulate the polytomy as a null hypothesis, in reality, I expect that in most cases the branch has positive branch length. Thus, I expect to reject the null often, in contrast to usual applications of the frequentist test. The analyst should specify in advance the branches for which a polytomy null hypothesis is reasonable. This adds subjectivity, but such problems are always encountered with frequentist tests, and mine is no exception. My test also suffers from all the various criticisms leveled against the frequentist hypothesis testing [9] and the interpretation has to avoid all the common pitfalls [81].

4.5 Conclusions

I presented a statistical test, implemented in ASTRAL, for the null hypothesis that a branch of a species tree is a polytomy given a set of gene trees. My test, which relies on the properties of the multi-species coalescent model, performed well on simulated and real data. As expected, its power was a function of branch length, the number of genes, and the gene tree estimation error.

4.6 Acknowledgements

Chapter 4, in full, contains material from SAYYARI, E., AND MIRARAB, S. Testing for polytomies in phylogenetic species trees using quartet frequencies. Genes 9, 3 (8 2018), 132. I was the primary investigator and author of this paper.

Chapter 5

Fragmentary gene sequences negatively impact gene tree and species tree reconstruction

Species tree reconstruction from genome-wide data is increasingly being attempted, in most cases using a two-step approach of first estimating individual gene trees and then summarizing them to obtain a species tree. The accuracy of this approach, which promises to account for gene tree discordance, depends on the quality of the inferred gene trees. At the same time, phylogenomic and phylotranscriptomic analyses typically use involved bioinformatics pipelines for data preparation. Errors and shortcomings resulting from these preprocessing steps may impact the species tree analyses at the other end of the pipeline. In this article, we first show that the presence of fragmentary data for some species in a gene alignment, as often seen on real data, can result in substantial deterioration of gene trees, and as a result, the species tree. We then investigate a simple filtering strategy where individual fragmentary sequences are removed from individual genes but the rest of the gene is retained. Both in simulations and by reanalyzing a large insect phylotranscriptomic data set, we show the effectiveness of this simple filtering strategy.

5.1 Introduction

Genome-scale reconstruction of species trees has become the standard practice in phylogenetics. A typical phylogenomic analysis starts by sequencing hundreds to thousands of loci using one of several sequencing strategies (e.g., transcriptomics, targeted amplicon sequencing, hybrid enrichment, etc.). Data from multiple loci may be then concatenated together to build a supermatrix, which is then analyzed using standard phylogenetic methods such as maximum likelihood (ML). The concatenation approach ignores potential discordance between gene trees and the species tree [147, 56], and has been proven statistically inconsistent [198] under the multi-species coalescent (MSC) model [178, 191]. An alternative approach, gaining in popularity, is to first estimate a gene tree for each locus (independently from other loci) and to then combine the gene trees using a summary method [60]. Several existing summary methods have been proven statistically consistent under the idealized conditions when gene trees are considered error-free; examples of summary methods used in practice include ASTRAL [164, 165], STAR [143], NJst/ASTRID [141, 255], and MP-EST [142]. While alternative approaches such as co-estimation [90, 139] and site-based MSC-based methods [38, 47] exist, these methods have been less frequently used, perhaps due to their computational requirements [280, 21].

Despite their growing application to real data [105, 264, 203, 189], the accuracy of summary methods is directly impacted by the accuracy of the input gene trees [160, 180, 229, 199, 274]. A well-studied source of gene tree estimation error (or uncertainty) is statistical noise due to lack of phylogenetic signal in short loci [161, 274]. This has motivated the development of methods for detecting and removing low signal genes [207, 274] or binning of loci to larger units [161, 20]. However, other factors, such as long branch attraction and missing data may also impact gene tree accuracy [229, 75], and these have been less thoroughly studied (but see [140, 61]).

The effect of missing data on the accuracy of single-locus or supermatrix tree reconstruction has been thoroughly studied [130, 268, 218, 184, 267]. In a summary method pipeline, missing data come in two forms, as previously noted by [97]. A species may be fully missing from some of the loci; I refer to this scenario (type I in [97]) as *missing tips* and to the patterns of presence/absence resulting from it as *taxon occupancy*. Alternatively, a species may be present with only partial data for some of the loci, and I refer to this scenario as *fragmentary data* (type II in [97]). These two forms of missing data may have very different impacts on the species tree reconstruction. Missing tips may negatively impact the summary method when the species tree is being inferred from a set of taxonomically incomplete (partial) gene trees, whereas, fragmentary data may negatively impact the gene tree inference step [97, 219]. While some studies have examined the impact of missing tips on summary methods [98, 272, 100], to my knowledge, only [97] have examined impacts of both types of missing data.

Current high-throughput genomic sequencing methods vary considerably in the size of the raw sequencing reads they generate, with there generally being a positive relationship between read length and the error rate in the sequence [170]. Sequence assembly (generating larger contigs) also varies greatly in efficiency and accuracy, and most applications, due to computational difficulty, rely on heuristics [131]. High-quality sequence generation and assembly also can be compromised by challenges with organism size and availability, particularly for highly diverse taxa of small body size such as insects [195]. Finally, transcriptomic datasets [264, 166] can have length variation in the assembled genes because of alternative splicing and the coverage of individual genes may also be affected by expression levels. Thus, phylogenomic and phylotranscriptomic studies that strive to provide thorough taxon sampling of diverse lineages at a reasonable cost often contain fragmentary data, at least for a subset

```
             Gene 1  Gene 2                        Gene 1  Gene 2
Species 1  AC----- ACCGATA            Species 1  ------- ACCGATA
Species 2  ACCGTTA ACCGATA     ⇒      Species 2  ACCGTTA ACCGATA
Species 3  CCCGTAA ACCGATA            Species 3  CCCGTAA ACCGATA
```

Figure 5.1: Example of filtering fragments from the concatenation alignment. Removing the fragmentary sequence Species 1 from Gene 1 only increases fragmentation without introducing any obvious benefit.

of the taxa. Since fragmentary data may negatively impact both the gene alignment [175] and gene tree estimation [130, 97], it is important to study effects of fragmentary data on the species tree reconstruction and ways to ameliorate the impacts.

One approach to deal with negative impacts of fragmentation, used in some phylotranscriptomic studies [264], is to remove each species from those genes where it is fragmentary. Filtering fragmentary data creates missing tips and thus presents a trade-off between fragmentation and taxon occupancy. Because of this trade-off, it is not clear whether filtering fragmentary data is overall beneficial to the accuracy; if indeed beneficial, it is not clear what level of filtering is warranted. Note that a similar trade-off does not face a concatenation analysis because no gene tree is ever estimated in such analyses and removing fragmentary sequences while keeping the respective genes only creates more missing data (Fig. 5.1) without any obvious benefit (except perhaps in the alignment step).

in this dissertation, I study effects of fragmentary data on species and gene tree reconstruction using summary methods. In line with observations of [97], but using simulations in addition to real data, I demonstrate the negative impact of fragmentary data. Unlike [97] who deal with fragmentary data by removing genes that show low phylogenetic signal, I study the strategy of filtering specific species from individual genes. Given a filtering threshold (e.g., 20%), I remove from each gene alignment any species that has non-gap characters in less than the given threshold (e.g., 20%). This form of filtering retains the gene and can arguably result in better utilization of the data because the non-fragmentary sequences are retained.

I test my proposed filtering method on an insect dataset and corroborate my findings in simulations.

I studied an empirical transcriptomic dataset of insects consisting of 1478 protein-coding genes of 144 taxa, where 27% of the alignment is gaps [166]. In 90% of the genes, there are 115 to 141 species present, and aligned protein coding sequences were between 134 and 890 amino acids in 90% of the genes.

Insects represent a species-rich lineage of organisms with generally small body size and can be challenging for production of high-quality phylogenomic data due to the difficulty of obtaining sufficient tissue (and thus DNA) for tiny, rare taxa required for full taxon representation. The insect lineage contains within its history several questions of broad evolutionary and scientific interest beyond its high species diversity, such as the evolution of wings and flight, various forms of metamorphosis, and multiple origins of social behavior [166]. Thus, for reconstructing the history of insects in sufficient detail to draw conclusions about biological questions of interest, a relatively full taxon sampling from all insect groups was desired. Transcriptomic data were used as an achievable way of providing such taxon representation with comparative genomic data for phylogeny estimation. Transcriptomic data typically include a large number of genes, but with the cost of having significant amounts of fragmentary data. [166] used concatenated analysis rather than coalescent-based gene tree summary methods, presumably because of the highly variable quality of the individual gene trees. The large size of this dataset and the high amount of fragmentary data make it especially well-suited for my analyses. In addition, the insect phylogeny has received considerable attention over the years, so that I have some prior expectation of relationships among some lineages, providing a perspective on the accuracy of the phylogenetic results.

In simulations, I study impacts of fragmentary data and the filtering strategy on the accuracy of gene trees and consequently the accuracy of the species trees. My simulated dataset (see Methods) simulates gene tree discordance due to ILS and I use estimated gene trees (with error); I also randomly inject fragmentation in gene alignments with patterns that emulate the biological insect dataset. I infer gene trees from the original sequences, unfiltered alignments, and alignments after filtering fragments (thresholds: 10% – 80%) using both RAxML [232] and FastTree2 [188]. I infer species trees using ASTRAL-II [165] using 50, 200, or 1000 gene trees.

5.2 Results

I start with a simulation study and then analyze the empirical insect dataset.

5.2.1 Simulation Results

Impact on gene trees

Comparing the original datasets that include no fragmentation (*orig-seq*) and those with injected fragmentation (*no-filtering*) shows that the presence of fragmentary data dramatically increases gene tree error (Fig. 5.2), as measured by the average normalized Robinson-Foulds distance (NRF) between true gene trees and the estimated gene trees. However, progressively applying more aggressive filtering gradually decreases gene tree error (Fig. 5.2) and the extent of improvements depends on the filtering threshold ($p \ll 10^{-5}$; ANOVA). With no filtering, the NRF distance is as high as 0.44 for FastTree and 0.39 for RAxML. Filtering fragmentary data gradually reduces the NRF distance. At the 75% threshold, the average NRF distance is reduced to 0.30 or 0.28, respectively, for FastTree and RAxML, which is only slightly lower than the average NRF error with no fragmentation (0.27 or 0.25, respectively) but these small differences are still statistically significant ($p \ll 10^{-5}$). Overall, RAxML gene trees are significantly more accurate than FastTree trees ($p \ll 10^{-5}$). Note that to compute the NRF distance, true and estimated trees are restricted to the same set of leaves, normalizing by the remaining branches; as Figure 5.3 demonstrates, reductions in the NRF after filtering cannot be attributed to the shrinking leaf set due to filtering.

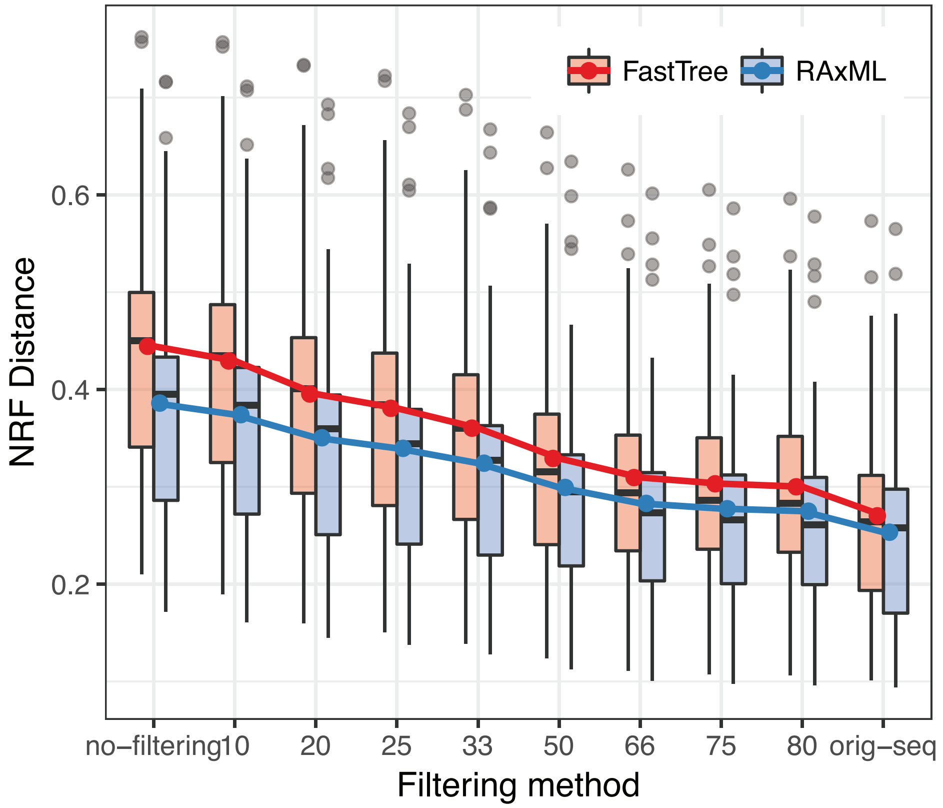


Figure 5.2: NRF distances between true and estimated gene trees in the simulated dataset. The x-axis shows the filtering thresholds; from left to right, more aggressive filtering is applied. Leftmost boxes are for no filtering and the rightmost boxes are for gene trees in absence of fragmentary data. For each threshold, box plots and average distances (red line) over 48 replicates are shown.

Impact on the species tree

Regardless of the number of genes, adding fragmentation to the data increases the NRF distance between true and estimated species trees (Fig. 5.4a) significantly ($p = 0.00014$ for RAxML and $p \ll 10^{-5}$ for FastTree). For example, with 1000 genes, the average species tree NRF distance increases from 0.030 with no fragmentation to 0.057 (90% increase) with fragments and no filtering for FastTree gene trees and from 0.023 to 0.037 (60% increase) for RAxML gene trees.

Filtering fragmentary data has non-monotonic impacts on the species tree error

Table 5.1: Average species tree error for RAxML and FastTree gene trees with different number of genes and using different fragmentary filtering thresholds for the simulated dataset.

	#Genes	no-filtering	10	20	25	33	50	66	75	80	orig-seq
FastTree	50gt	0.105	0.107	0.101	**0.098**	0.101	0.109	0.130	0.142	0.150	0.079
FastTree	200gt	0.072	0.064	0.059	0.060	**0.058**	0.062	0.069	0.073	0.077	0.050
FastTree	1000gt	0.057	0.052	0.044	0.041	**0.036**	**0.036**	0.040	0.043	0.043	0.030
RAxML	50gt	**0.086**	0.087	0.088	0.086	0.090	0.107	0.126	0.134	0.143	0.069
RAxML	200gt	0.052	0.053	0.047	**0.046**	0.048	0.054	0.064	0.067	0.073	0.039
RAxML	1000gt	0.037	0.033	0.028	0.029	**0.026**	0.028	0.029	0.031	0.032	0.023

(Fig. 5.4a). As the filtering threshold increases, the average species tree estimation error initially tends to drop but eventually starts to increase again. The optimal threshold depends on the number of genes and in most conditions varies between 25% and 33% (Table 5.1). Limiting myself to all thresholds up to 33%, I observe that the accuracy of the species tree tends to gradually improve as a result of increased filtering when at least 200 genes are available; however, improvements are statistically significant only with 1000 genes ($p = 0.00461$ and $p = 0.0417$, respectively for FastTree and RAxML) and not for 200 ($p = 0.102$, and $p = 0.314$, for FastTree and RAxML). With 1000 genes, filtering is never worse than no filtering, even with extremely aggressive filtering. At the 50% threshold, the NRF distance reduces from 0.037 to 0.028 for RAxML gene trees and from 0.057 to 0.036 for FastTree trees.

Gene tree versus species tree error

Reducing gene tree error by increased filtering is only beneficial to the species tree estimation when taxon occupancy is not dramatically sacrificed (Fig. 5.4b). As I go from no filtering to filtering up to 33%, the species tree error and gene tree error both tend to decrease at first. Further increases in the filtering threshold continue to reduce the gene tree error, but those reductions don't always translate to improvements in species tree error, and in fact, can increase it. This is perhaps partly because improvements in gene tree error eventually become small with each increase in the filtering threshold. More importantly, the taxon occupancy continues to decrease with more filtering and lack of occupancy may offset the benefits of reduced gene tree error. The average taxon occupancy drops from 88% to 77% and then to 65% as I increase the filtering threshold from 33% to 50% and then to 66%.

Table 5.2: Significant clades in the insect phylogeny with references to evidence supporting them. Letter codes refer to red-labeled nodes in Figure 5.8, except node I that is missing in my final species tree but present in the [166] final concatenation tree. The evidence from the literature fall into three groups. *Strong*: virtually always recovered in previous phylogenetic studies and not controversial based on comparative morphology *Fairly strong*: often recovered by phylogenomic studies but either not clearly supported by morphology or sometimes another well-supported alternative exists. *Weak:* either controversial based on comparative morphology or seldom strongly supported by any analysis.

Code	Clade composition	Evidence	Selected references
A	Mecoptera+Siphonaptera	Fairly strong	[266]
B	Diptera+ (Mecoptera+Siphonaptera)	Strong	[119, 266, 24]
C	Trichoptera+Lepidoptera	Strong	[120, 117]
B/C	Clades B+C	Strong	[266, 26, 25]
D	Neuropterida+ (Coleoptera+Strepsiptera)	Fairly strong	[266, 24, 176, 33]
D/B/C	Holometabola minus Hymenoptera	Fairly strong	[209, 167, 155, 92]
E	Holometabola (Endopterygota)	Strong	[266, 24, 176, 252]
F	Hemiptera+Thysanoptera	Fairly strong	[25]
G	Acercaria+Hymenoptera	Fairly strong	[25]
H	Mantophasmatodea+ Grylloblattodea	Fairly strong	[250, 42, 269]
I	Clade H+(Embiidina+Phasmida)	Weak	[166] but not in the final species tree of this study
J	Polyneoptera (Orthopteroidea)	Fairly strong	[275, 102, 132]
K	Neoptera	Strong	[117, 113]
L	Pterygota (winged insects)	Strong	[118, 83, 88, 113, 262]
M	Zygentoma+Pteygota	Fairly strong	[28, 63, 113]
N	Insecta	Strong	[28, 63, 113, 91]
P	Psocodea+Holometabola	Weak	Only supported by [166] and weakly by this study
Ingroup	Hexapoda	Fairly strong	[112, 113, 84, 158, 194]

5.2.2 Empirical Results

Reconstructed ASTRAL-II insect trees (Figs. 5.10– 5.17) change in topology and support based on my choice of gene tree estimation method (RAxML versus FastTree) and the treatment of fragmentary data. Before presenting results in detail, I start by describing my approach in judging accuracy on the biological dataset.

Evaluation and expected relations

In order to gain some indication of how accurate the species tree results are, I first surveyed some relationships that have been previously considered to be well established on the basis of evidence, and also identified those that have been previously found but not consistently supported (Table 5.2). Among the former are the monophyly of Hexapoda among the Pancrustacea, monophoply of "true insects", monophyly of the Dicondylia (Zygentoma + Pterygota), monophyly of Pterygota, monophyly of Neoptera, monophyly of Holometabola, monophyly of Antliophora (Diptera + Mecoptera + Siphonaptera) and monophyly of Amphiesmenoptera (Trichoptera + Lepidoptera). Among the latter are the monophyly of the Paleoptera (Ephemeroptera + Odonata), monophyly of the Polyneoptera, sister groups relationships between the Grylloblattodea and Mantophasmatodea, between Hemiptera and Thysanoptera, and between Coleoptera and Strepsiptera, and the position of Hymenoptera as sister to the remaining holometabolous orders. [166] found two other relationships without previous strong support: the sister group relationship between Psocodea and Holometabola, and a clade containing Mantophasmatodea + Grylloblattodea plus Embiidina + Phasmida. I discuss my results in the context of these prior expectations and findings.

In addition to judging the quality of the species tree based on prior evidence, I also study the impact of the filtering on taxon occupancy, gene tree bootstrap support, and evolutionary diameter of the gene trees, measured by the tip-to-tip distance. A reduced taxon occupancy is clearly undesirable (even if inevitable). Reduced gene tree bootstrap support can be interpreted as a sign of increased uncertainty about gene trees and perhaps increased error. An increase in the evolutionary distance can be indicative of artificially long branches that can be inferred as a result of fragmentary data [130].

Occupancy

Filtering fragmentary data affects the taxon occupancy of different orders and species unevenly (Fig. 5.6a and Fig. 5.5). Here, I measure occupancy of a clade by the percent of genes that have at least one of the species from the specific clade. Almost all clades have at least 50% occupancy, regardless of the threshold selected. However, the occupancy of clades for the filtering thresholds of 20% to 33% are similar, with a considerable drop at 50%, and a dramatic drop at 66% filtering or higher. At 50% filtering, the occupancy is above 70% for all orders.

Gene trees

Filtering more fragmentary data improves gene tree bootstrap support for both RAxML and FastTree (Fig. 5.6b). For example, with no filtering, the number of branches with 100% support in RAxML gene trees is only 5% but gradually increases to 12% with the highest level of filtering (i.e., 80% threshold). Similarly, the median branch bootstrap support for RAxML (or FastTree) gene trees is 37% (29%) with no filtering but gradually increases to 48% (48%) with 80% filtering. Conversely, the number of branches with less than 33% support decreases from 47% (53%) with no filtering to 38% (39%) with 80% filtering (Fig. 5.6b). Overall, both RAxML and FastTree gene trees improve in their bootstrap support, but the improvements are larger for FastTree.

The increased bootstrap with increased filtering can be attributed to the negative impact of fragmentary sequences on estimated gene trees (both ML and bootstrap replicates). Fragmentary sequences and their treatment as ambiguous data tend to result in long branches in an ML estimation [130]. Consistent with this expectation, I observe that filtering out fragmentary data consistently reduces the evolutionary diameter (Figs. 5.6c and 5.7) of gene trees, indicating that fragments may result in long branches. This reduction in evolutionary diameter coincides with increasing gene tree bootstrap support (Fig. 5.6c). In the case of bootstrap replicates used to estimate support, the problem of fragmentation is exacerbated by the resampling of sites, which may leave no or few sites with any non-gap characters. Finally, note that the perceived improvements in gene trees support and evolutionary diameter come at the cost of reduced taxon occupancy (Figs. 5.6c).

Overall, the occupancy plot (Fig. 5.6a), gene tree statistics (Fig. 5.6bc), and my simulations (Fig. 5.4a) lead us to favor thresholds between 20% and 50%. In the rest of the discussion, while I will continue to discuss all thresholds, where I need only one threshold, I will use 50% as the default (I return to this choice in the discussion section).

Species trees

With all fragments included, the ASTRAL tree inferred from FastTree gene trees fails to recover a number of relationships (Fig. 5.8a): the monophyly of Hexapoda (ingroup), monophyly of Dicondylia (M), monophyly of Pterygota (L), monophyly of Neoptera (K), monophyly of Thysanoptera + Hemiptera + Psocodea + Holometabola (G), and monophyly of Neuropteroidea (D). As filtering of fragments increases (moving left to right in Fig. 5.8a), the species tree improves, so that monophyly of Hexapoda (Ingroup), Dicondylia (M), Neoptera (K), and Thysanoptera + Hemiptera + Psocodea + Holometabola (G) are recovered, albeit not strongly so in the case of Neoptera and Hexapoda. ASTRAL run using FastTree gene trees never recovers the monophyly of Pterygota (L) and monophyly of Neuropteroidea (D), which have fairly strong support in the literature (Table 5.2). Neither does it recover the Psocodea + Holometabola clade (P) or the I node found by [166], which don't have strong support in prior analyses. Some expected clades, especially within Holometabola, always had strong support no matter how much filtering was done (Fig. 5.9).

In contrast, when RAxML gene trees are used, the only strong relationship that is not recovered is the monophyly of Hexapoda (Fig. 5.8ab). Any level of filtering would result in a monophyletic Hexapoda, showing that even for RAxML gene trees, correct handling of fragmentary data can improve the species tree topology. Moreover, even though the species tree topology inferred from RAxML gene trees is relatively robust in the presence of fragmentary data, the ASTRAL estimated branch lengths increase with filtered gene trees (Fig. 5.6d). Since coalescent unit branch lengths tend to be underestimated [210, 161], the increased branch lengths are likely to be more accurate. I note that my simulations corroborate that coalescent unit branch lengths are under-estimated and show a strong positive correlation between gene tree accuracy and estimated branch lengths (Fig. 5.18).

My final ASTRAL tree (Fig. 5.8b) using RAxML gene trees and 50% filtering includes all major clades with prior support in the literature, and all but two of them (monophyly of Hexapoda, and Psocodea + Holometabola) had full support. Interestingly, [166] found a clade of Embiidina, Phasmida, Grylloblattodea and Mantophasmatodea (I) that my final tree does not recover, but this clade has little historical support; it will be interesting to see if this clade is supported by further studies.

Gene tree species tree discordance

As a final evidence that gene tree accuracy has improved, I demonstrate that after filtering, both RAxML and FastTree gene trees show reduced discordance. Overall, the amount of gene tree discordance with the species tree reduces substantially as I increase filtering, especially up to the 66% threshold (Fig. 5.6e). These reductions are the reason for the reduced coalescent unit branch lengths (Fig. 5.6d). Similarly, the overall ASTRAL quartet score (proportion of gene tree quartets found in the species tree) increases as more filtering is applied (Table 5.3).

To further break down patterns of discordance, we compare support for major clades (orders plus clades shown in Table 5.2) in my gene trees before and after filtering (Fig 5.19). Before filtering, many of the insect orders receive surprisingly little support in my gene trees.

Table 5.3: ASTRAL quartet scores on the biological insect dataset for RAxML and FastTree gene trees using different filtering thresholds.

	FastTree	RAxML
20	0.6742	0.7173
25	0.6795	0.7190
33	0.6828	**0.7194**
50	0.6899	0.7159
66	0**.6914**	0.7091
75	0.6893	0.7044
80	0.6886	0.7014
no-filtering	0.6625	0.7131

For example, with RAxML (or FastTree) gene trees, 13 (or 14) out of 26 orders are recovered in less than half of the gene trees, and only 6 orders are recovered in at least three-quarters of the gene trees. Moreover, most gene trees have low support and cannot strongly reject or support the monophyly of these orders. However, the 50%-filtered gene trees show strong support for most orders. Only 7 orders are recovered in fewer than half of these gene trees, and the number of orders supported by at least three-quarter of the genes increases to 13 and 11, respectively for RAxML and FastTree. As an example, before filtering, 35% (28%) of RAxML (FastTree) genes recovered Lepidoptera as a monophyletic clade, and only 15% (9%) have high bootstrap support, whereas, after filtering, 64% (62%) recover it and 47% (44%) have high support. While Lepidoptera has one of the biggest changes, patterns across all orders consistently point to reduced discordance. Major clades (other than orders) also have increased support and substantial reduction in highly supported discordance.

5.3 Discussion

5.3.1 Impacts of fragments

I showed that fragmentary data can have substantial negative impacts on gene trees and consequently species trees estimated in a summary method pipeline. These results build on previous studies of weak phylogenetic signal, the resulting high gene tree error (or uncertainty), and species tree error [161, 180, 207, 274]. It is important to note that fragmentation not only weakens the signal, but it may also create biases in an ML analysis [130].

The harmful impact of fragmentation was previously observed by [97]. My results corroborate their observation. However, I propose a very different solution. Unlike their solution of removing the entire gene, which can lead to loss of otherwise useful signal and perhaps a non-random sample of genes [99], I propose removing specific problematic taxa. Importantly, I observed that trading off decreased taxon occupancy with decreased levels of fragmentation (type I versus type II in the terminology of [97]) is beneficial, but only to a point; excessive filtering can also impact the accuracy of the species tree by creating missing data in gene trees. The amount of improvement depended on the number of genes, and filtering did not seem useful when only a small number of genes was available. The reductions in gene tree error were substantial (e.g., from 0.39 to 0.28); improvements in species tree topological accuracy may be considered small in magnitude (0.01–0.02 NRF), but I note that the error is reduced by a quarter of the original error, and that, these improvements come at no extra cost. I further note the improvements in species tree branch lengths.

Consistent with the literature [98, 272, 100], my results indicate that summary methods are somewhat robust to missing data, but I also show that this robustness has limits as seen by [97]. In the context of a single maximum likelihood analysis, [267] observed that the absence of enough data, and not the presence of missing data *per se*, can cause inaccuracy. Importantly, I do not filter entire genes because they miss some taxa or because they have some fragmentary sequences. Prior research suggests aggressive filtering of entire genes with missing data can be harmful [99, 236]. My results do not conflict with those studies and my filtering approach is in fact motivated by their observations.

Finally, I observe that the number of genes has the strongest effect on the species tree error. Therefore, removing genes is not desirable. Instead, when possible, increasing the number of genes may improve the species tree topological accuracy even in the presence of fragments.

Filtering threshold

The best choice of the filtering thresholds will always depend on the dataset. However, my analyses suggest a possible way forward for systematists. Since ML tools such as FastTree can easily compute many hundreds of gene trees quickly and relatively accurately, one can examine different thresholds empirically. By changing the threshold, reestimating gene trees, and computing occupancy (q), average gene tree support (p), and evolutionary diameter (d) (e.g., as in Figs 5.6bc), analysts can look for thresholds that reduce occupancy minimally while increasing support or decreasing long branches substantially. This involves making trade-offs and the best way of making such trade-offs requires further analyses.

Some simple rule-of-thumbs could be designed, and Table 5.4 gives several rules and applies them to the insect datasets. For example, I can simply use the threshold that maximizes pq/d. Another simple rule is using pq, which would emphasize occupancy and support equally. Based on the belief, backed by the literature, that that reduced occupancy is less damaging than high gene tree error, I can use p^2q to weight bootstrap support more than occupancy. Finally, one can pick the threshold that gives the highest p given that q is above a threshold (say, 70%). On the insect dataset, these rules selected thresholds between 20%

Table 5.4: Rule-of-thumb metrics to find the best filtering threshold, applied to the insect dataset. p indicates average gene tree bootstrap supports (over all genes), q indicates average occupancy over genes, and d indicates evolutionary diameter (here average of average gene tree branch lengths). The best threshold for each metric is shown in bold.

Threshold	Method	pq	p^2q	pq/d	$pq/(1-p)$	p if $q > 0.7$	d	p	q
no-filtering	RAxML	0.3866	0.1635	0.3068	0.6700	0.4231	1.2600	0.4231	0.9136
20	RAxML	**0.3936**	0.1772	0.3212	0.7157	0.4501	1.2253	0.4501	0.8744
25	RAxML	0.3921	0.1802	0.3227	0.7254	0.4595	1.2150	0.4595	0.8533
33	RAxML	0.3877	0.1843	**0.3247**	0.7388	0.4752	1.1941	0.4752	0.8158
50	RAxML	0.3653	**0.1847**	0.3177	**0.7390**	**0.5057**	1.1499	0.5057	0.7223
66	RAxML	0.3306	0.1755	0.2993	0.7047	NA	1.1043	0.5309	0.6226
75	RAxML	0.3079	0.1671	0.2859	0.6733	NA	1.0770	0.5427	0.5673
80	RAxML	0.2953	0.1625	0.2796	0.6566	NA	1.0562	0.5503	0.5366
no-filtering	FastTree	0.3430	0.1288	0.3432	0.5492	0.3754	0.9995	0.3754	0.9136
20	FastTree	0.3519	0.1416	0.3569	0.5889	0.4024	0.9860	0.4024	0.8744
25	FastTree	0.3549	0.1476	0.3630	0.6075	0.4159	0.9776	0.4159	0.8533
33	FastTree	**0.3569**	0.1562	0.3701	0.6346	0.4375	0.9644	0.4375	0.8158
50	FastTree	0.3475	0.1672	**0.3724**	0.6697	**0.4811**	0.9332	0.4811	0.7223
66	FastTree	0.3236	**0.1682**	0.3597	**0.6739**	NA	0.8997	0.5198	0.6226
75	FastTree	0.3047	0.1636	0.3466	0.6581	NA	0.8790	0.5370	0.5673
80	FastTree	0.2932	0.1602	0.3393	0.6464	NA	0.8641	0.5464	0.5366

and 50% for RAxML gene trees (Table 5.4). Since the 50% threshold was chosen the most often, I chose to use it as my default threshold. On the simulated dataset, for FastTree gene trees, where I could perform bootstrapping, several of my rules (e.g., p^2q and $\frac{pq}{1-p}$) tend to select thresholds between 20% and 50% (Fig. 5.21); these match the optimal thresholds for FastTree simulations (Fig. 5.4).

Even though these rules seem to pick reasonable thresholds on my insect data, whether any of them performs well on a wide range of datasets remains unclear and require future studies. Moreover, filtering in general, and the use of bootstrap support in particular, could always add a bias, and thus, I do not suggest using thresholds that are varied from gene to gene.

Finally, if some relationships are judged very strong by prior evidence, detecting whether one recovers them may also prove useful, though this strategy should be used judiciously to avoid confirmation bias. In absence of extensive analysis, thresholds 25% – 50% seemed reasonable in my simulated and empirical analyses and may prove useful as a default for other analyses.

RAxML versus FastTree

While the choice of the ML method for inferring gene trees was not the focus of my study, my simulation analyses showed a clear advantage in using RAxML versus FastTree. Moreover, on the biological dataset, using RAxML (with automatic model selection) rather than FastTree (with a fixed model) led to further improvements in the species tree, recovering the monophyly of Neuropteroidea (node D) and Pterygota (node L), both of which have strong evidence from the literature. It is also interesting that the support for Neoptera (Node K) and the tentative sister-group relationship of Psocodea and Holometabola found by Misof et al.'s concatenation results (node P) were increased using RAxML gene trees. Overall, simulations and real data indicate that not only gene trees are less accurate when estimated using FastTree, but also, the ASTRAL species trees inferred from FastTree gene trees are less accurate than those inferred from RAxML gene trees. Interestingly, FastTree gene trees consistently have reduced branch lengths compared to RAxML trees on the biological dataset (Fig. 5.6c), perhaps because of FastTree's extensive use of the minimum evolution criteria (in addition to maximum likelihood). Finally, I note that FastTree does not allow for extensive model selection (for proteins), a fact that on biological (but not simulated) datasets could further contribute to its inaccuracies.

A previous independent simulation study by [138] had concluded that the two methods are essentially identical in terms of accuracy. The opposing conclusion drawn by [138] and my study may be related to simulation conditions. My study considers conditions that include short branches prone to ILS but includes no alignment error; in contrast, [138] use datasets originally simulated to study alignment accuracy and include very divergent sequences (at least 50% average p-distance between sequences). To my knowledge, mine is the first simulation study to show that RAxML gene trees are more accurate, and I believe, the results should discourage analyses that rely solely on FastTree. While many practitioners have perhaps already suspected that the much slower RAxML algorithm is more accurate under some conditions, the results shown here provide direct comparative evidence.

Despite the difference in accuracy, impacts of fragmentation had broadly similar patterns, regardless of the gene tree method used. Therefore, I believe for exploratory analyses of a dataset, the use of FastTree is justified whereas final analyses used to infer the species tree are more reliable when based on the RAxML gene trees. Future work should test if using Bayesian methods for estimating gene trees would similarly improve the species tree accuracy.

Insect phylogeny

As the debate between concatenation and summary methods pipelines continues [61, 220, 228, 273], I note that [166] had only used concatenation in their analyses. My final ASTRAL tree using RAxML gene trees is highly congruent with the concatenation tree of [166]. This result has several implications for insect phylogeny. Overall, it supports nearly all of the results of their concatenation analysis with respect to the major events in insect evolution, using a different analysis strategy. Where their final tree showed relationships with weak support, generally mine did as well, indicating that some results may require further effort to resolve with confidence. It is likely that with the generation of phylogenomic data with lower fragmentation (e.g., using full genomes instead of transcriptomes), gene tree summary methods will be able to improve upon the results of both studies.

5.3.2 Methodological limitations

It is important to note that even in the final RAxML gene trees, extensive gene tree discordance remains, and some of the discordance is highly supported (Fig 5.19). The presence of highly supported discordance can in principle favor summary methods over concatenation. However, I note that my analyses that included fragments produced results that strongly conflicted with strong evidence from the prior literature. Thus, the choice is not only between concatenation and summary methods but more broadly about choosing data generation methods and tailoring the analysis pipeline to the data. Summary methods can only produce good results when provided with good gene trees and removing fragmentary data and other sources of the error are essential to that goal. In this chapter, I demonstrated negative impacts of fragmentary data and sub-optimal gene tree estimation methods. However, several other sources of error were not addressed.

Even after filtering, the proportion of genes that fail to recover major insect orders remains arguably high. It is likely that gene tree error persists even after filtering. One major cause of the remaining discordance is likely the lack of strong signal in gene trees. In addition to insufficient signal, my models of sequence evolution are likely violated in many ways, especially when I consider 400+ million years of evolution, as I did here. Factors that include convergent effects of strong selection or unexpectedly high sensitivity to individual sites [217] may lead to systematic biases. Finally, even when the gene tree discordance is real, it may be due to factors other than ILS, including incorrect detection of orthology. Future work should explore improved scalable methods of dealing with these difficulties.

5.4 Materials and Methods

5.4.1 Analysis pipeline for insect dataset

I used the amino acid sequence data provided by [166] as “Supplementary 7”.

Filtering strategy: In real data, gaps can appear for two reasons: insertions and deletions (as inferred by an alignment algorithm) and missing data. My goal is to filter out sequences that are fragmentary (partially sequenced or assembled) but I don’t wish to remove sequences due only to indels. Defining what is a fragmentary sequence is complicated by the presence of gappy sites. Also, very gappy sites increase running time but provide little signal to the maximum likelihood analyses, which treat them as missing data (and not as indel signal). To address both issues, before identifying fragmentary sequences, I first remove extremely gappy sites, defined as those with more than 90% gaps. While this filtering can remove photogenically informative indels, I note that indels are not incorporated in models of sequence evolution used in my gene tree estimation tools. I then remove species that have less than 20% (1/5), 25% (1/4), 33% (1/3), 50% (1/2), 66% (2/3), 75% (3/4), or 80% (4/5) amino-acids (i.e., characters other than gaps). In order to filter sequences, I use a tool called seqtools, implemented as part of the PASTA [163] package. After filtering sequence, I re-estimate gene trees, but I keep the same alignment. In order to track the occupancy and bootstrap support, I use in-house scripts, available online `https://github.com/esayyari/discoVista`.

Gene trees: After each round of filtering, gene trees are estimated using FastTree2 [188] using its default amino acid substitution model, which is JTT [108] or RAxML [232] with the automatic amino acid model selection.

To infer my bestML gene trees, I use RAxML [232], version 8.2.9 with 10 runs of inference using different starting trees.

Table 5.5: The number of genes estimated with various models computed using RAxML automatic model selection.

Model	BLOSUM62F	DAYHOFF	DAYHOFFF	JTT	JTTDCMUT	JTTDCMUTF	
#Genes	1	5	8	154	47	30	
Model	JTTF	LG	LGF	VT	VTF	WAG	WAGF
#Genes	156	643	379	20	17	8	10

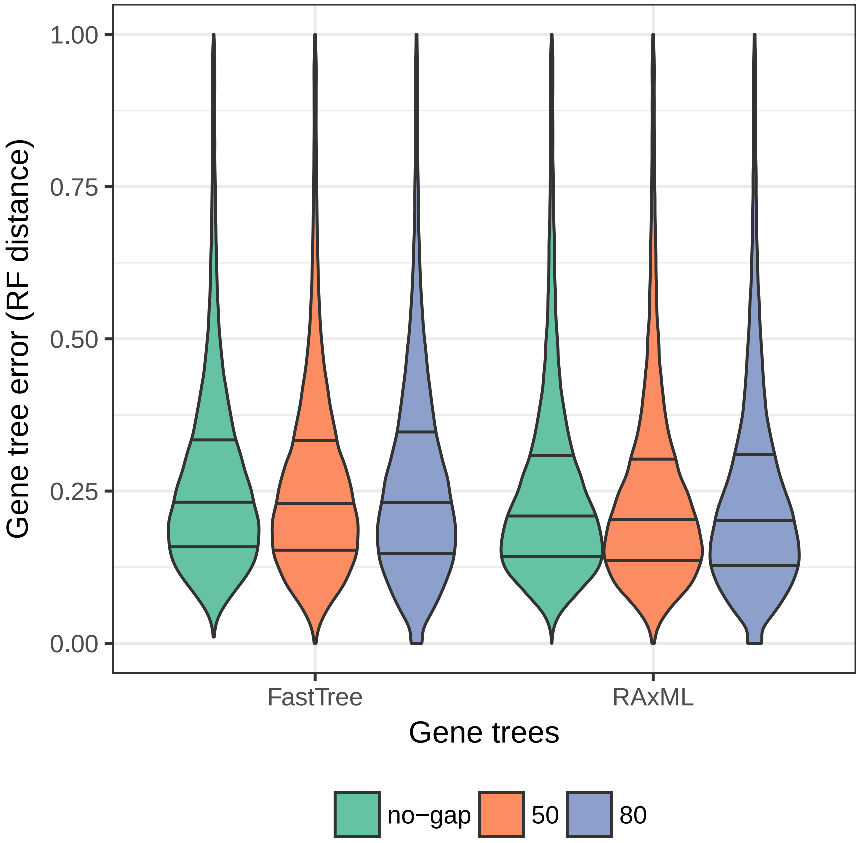


Figure 5.3: Impacts of using the NRF distance. The distribution of the NRF distance is shown for four sets of gene trees (two for RAxML and two for FastTree). “no-gap”: shows the NRF of the original estimated gene trees before addition of any fragmentations and computed on the full set of 101 taxa. “50” and “80”: I restrict the leafset of the *original* gene trees to be the same as the leafset of gene trees estimated on filtered gene sequences with a 50% (orange) or 80% (purple) filtering threshold. If NRF was sensitive to taxa present or if reductions in NRF after filtering of fragments were due to the removal of difficult edges and/or introduction of long branches, I would expect that the *restricted* case should have lower NRF than the *original*. Instead, I see very similar NRF between full trees and the trees on taxon sets that match the filtered gene trees. I conclude improvements in gene tree accuracy in Figure 5.2 are not related to the way NRF distance is defined and computed.

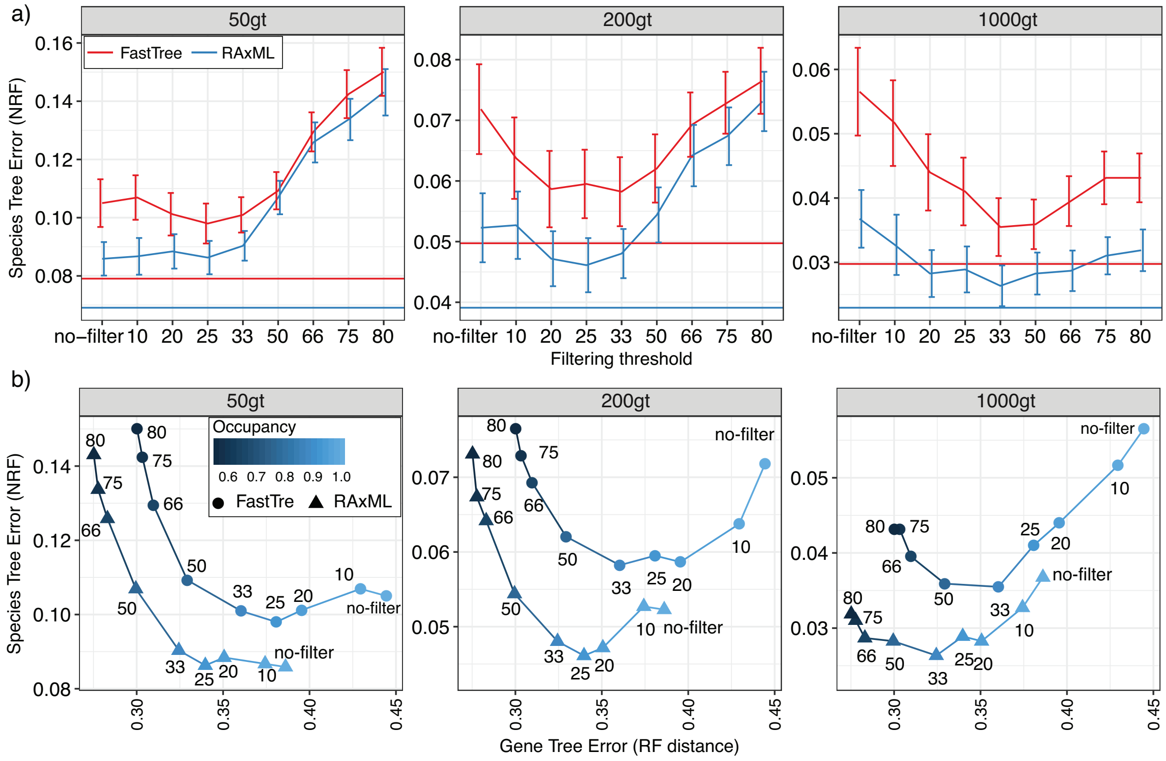


Figure 5.4: Species tree error in simulation datasets. a) NRF error of estimated ASTRAL species trees for different numbers of genes (boxes) and varying filtering thresholds (x-axis) with both RAxML and FastTree gene trees. The horizontal lines indicates the error rate of ASTRAL in the absence of fragmentary data. The error bars in these figures indicate the standard errors around the average. b) Correlation between the gene tree and the species tree error. The y-axis shows the average species tree error (NRF distance) and x-axis shows the NRF distance between true and estimated gene trees. Color shades represent the average occupancy of species, and fragmentary filtering thresholds are noted next to the dots. Results from RAxML and FastTree gene trees are distinguished by dot shape.

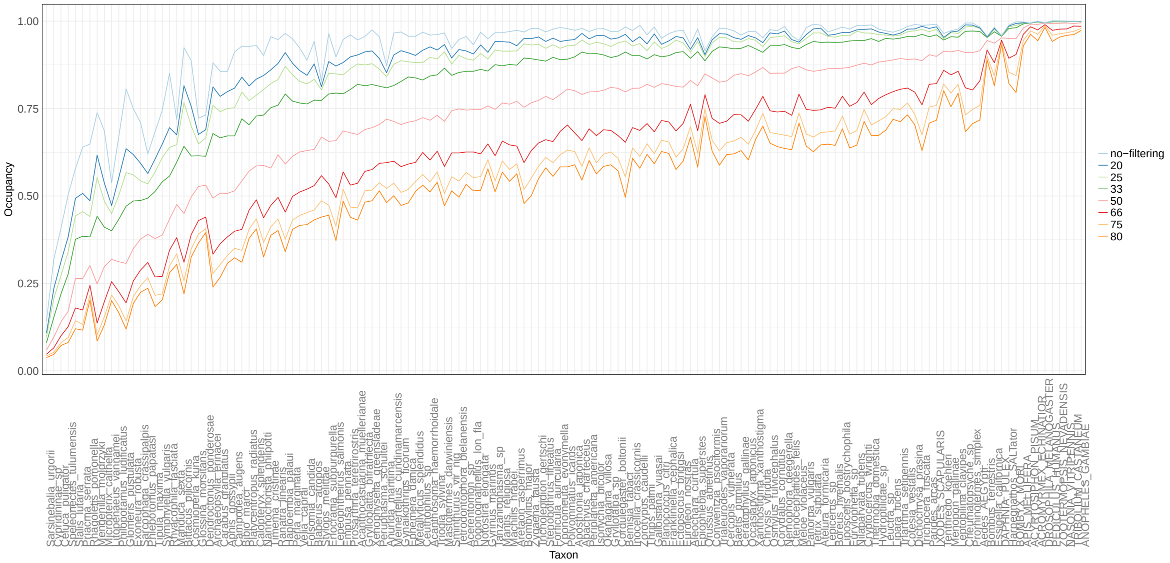


Figure 5.5: Occupancy of species in insects dataset. This figure shows the occupancy of species for insects dataset when different filtering thresholds applied. For e.g. filtering threshold of 50 means, species that has less than 50 percent of their sequence length as non-fragmentary data will be removed before estimating gene trees. With bigger thresholds, more species will be filtered out, and the species have lower occupancy.

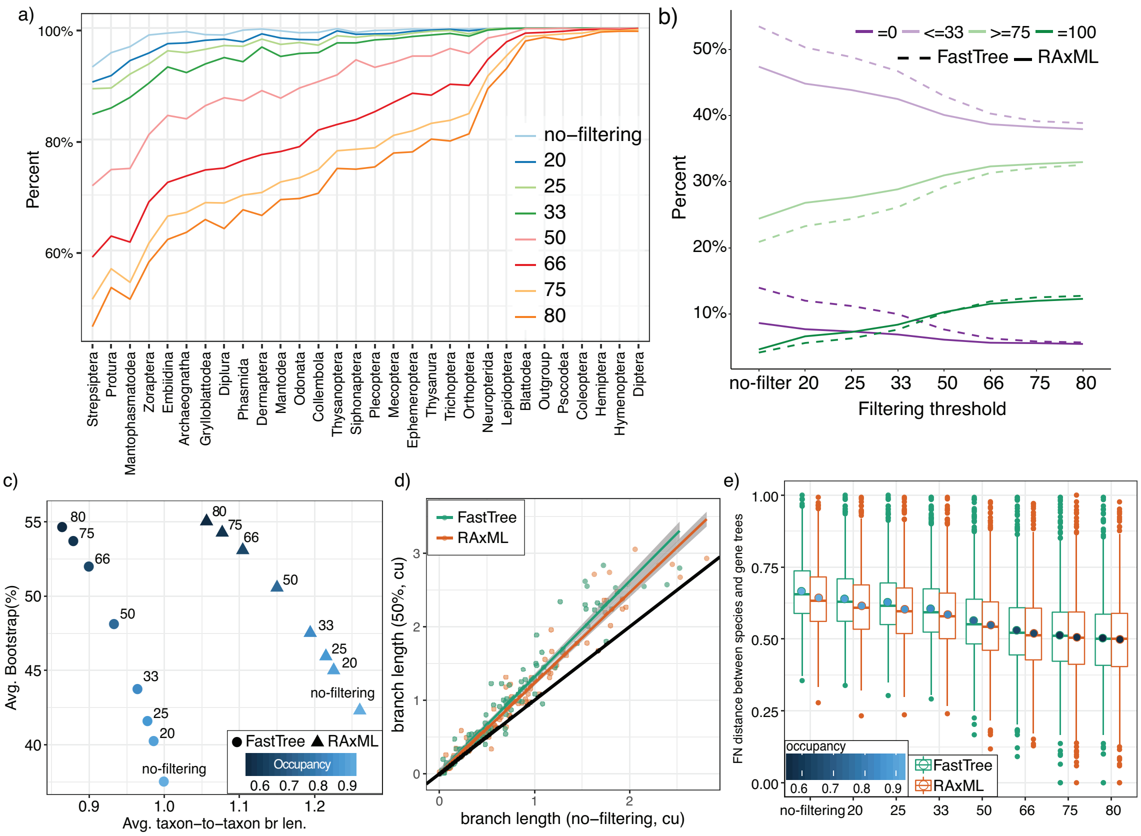


Figure 5.6: Impacts of filtering on the biological insect dataset. a) Occupancy of major clades after filtering fragmentary data with various thresholds (colors). b) Distribution of average BS values for different filtering thresholds. I show the percent of branches that have BS value of 0% (dark purple), less than 33%, (light purple), more than 75% (light green), and 100% (dark green). c) Average gene tree bootstrap support (y-axis) versus average (over genes) of average (over leaves) of root-to-tip distances (x-axis) with different filtering thresholds (text next to the dots). I use color to represent the average occupancy, where darker colors represent lower average occupancies. Average of maximum tip-to-tip distances (evolutionary diameter) shows similar patterns (Fig. 5.7). d) Coalescent unit branch lengths computed by ASTRAL. Each dot corresponds to a branch and its coalescent unit length is shown when estimated from unfiltered gene trees (x-axis) or 50% filtered gene trees (y-axis). Several branches (25 for FastTree and 13 RAxML) that were not shared between the two trees are removed. A line is fitted to all the points corresponding to each method, and the dashed line shows the unity line. e) Discordance of gene trees with various filtering thresholds (x-axis) versus the corresponding ASTRAL species tree. Boxplots show distributions of the proportion of species tree branches not found in gene trees.

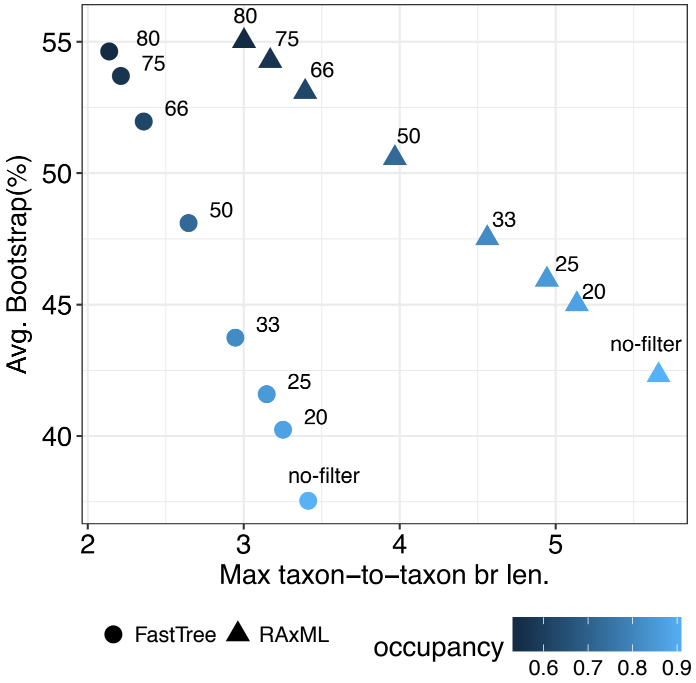


Figure 5.7: Impacts of filtering on the evolutionary diameter of gene trees in Insects dataset. Average gene tree bootstrap support (y-axis) versus average of maximum root-to-tip distances (x-axis) with different filtering thresholds (text next to the dots) for RAxML and FastTree gene trees. I use color to represent the average occupancy, where darker colors represent lower average occupancies.

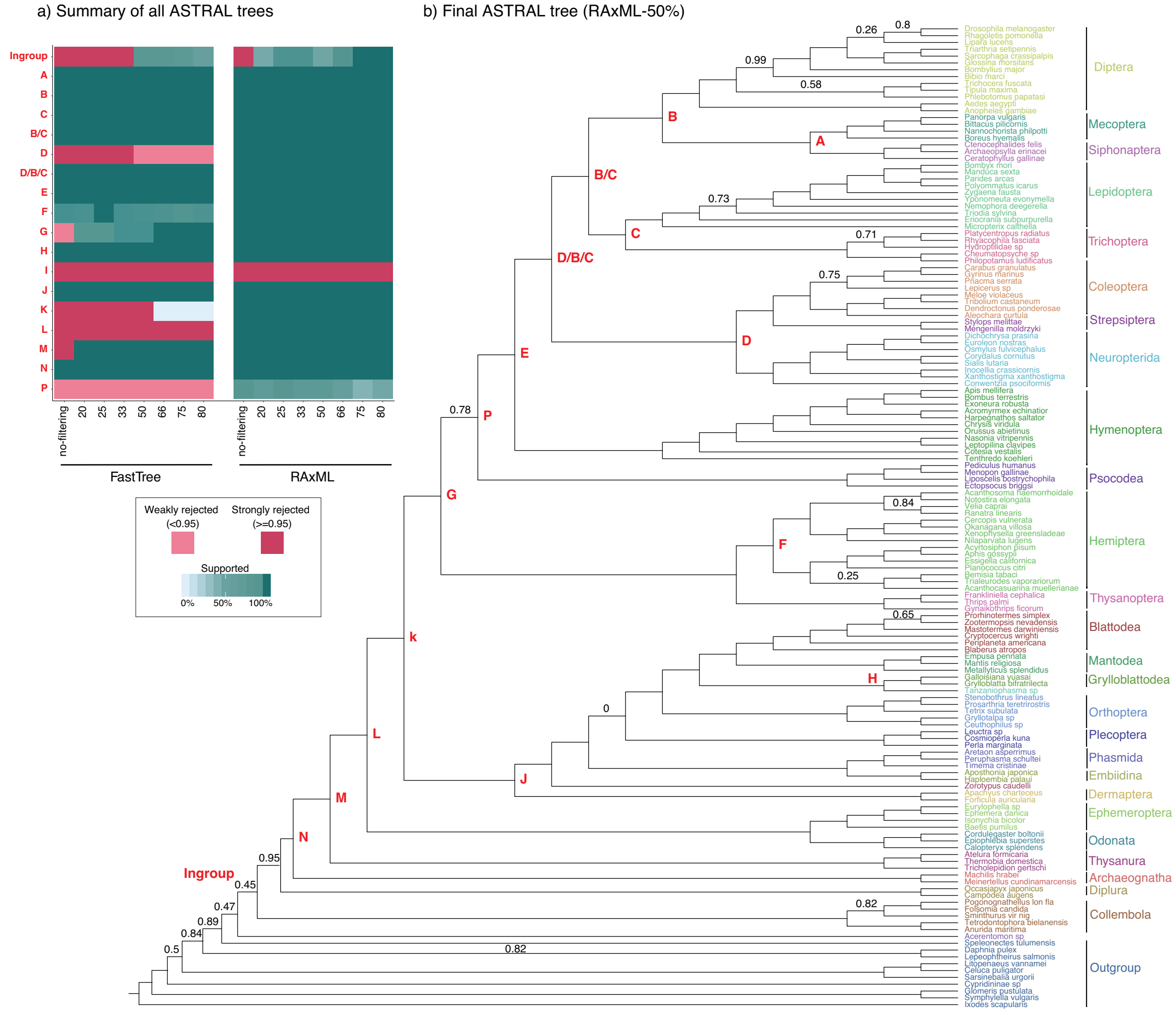


Figure 5.8: ASTRAL species trees. (a) Recovery of important clades in ASTRAL species trees with various filtering thresholds (see Figs. 5.10 – 5.17 for trees), represented by columns. Rows show important clades A to P (Table 5.2). The blue-green indicates monophyly of a clade, and the spectrum of blue to green colors show ASTRAL localPP support values [210]. Red, indicates strong or weak rejection of a clade. Weak rejection is defined as a clade that is absent from a tree but is compatible with the tree if branches below 95% support are contracted. See Figure 5.9 for more clades. (b) The ASTRAL species tree using 1478 RAxML gene trees with the 50% filtering threshold.

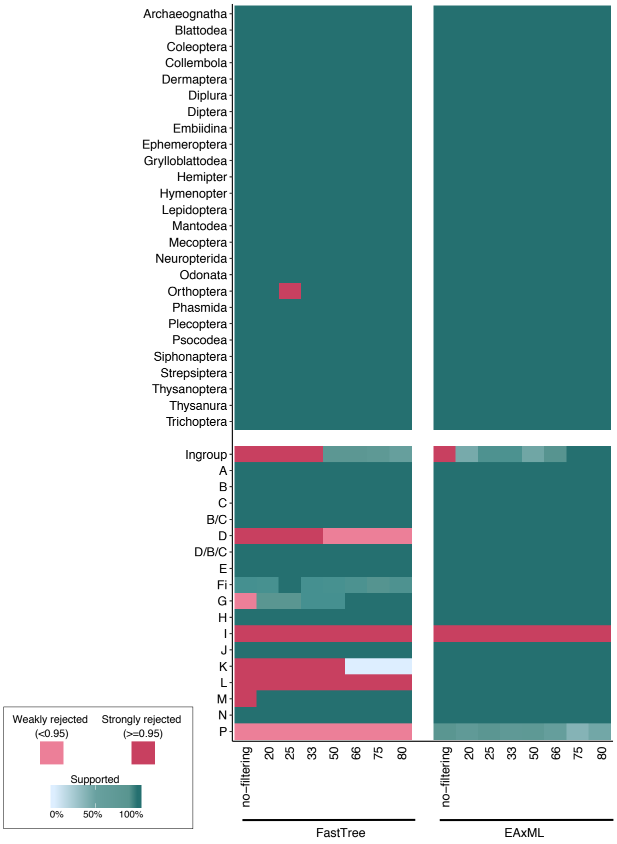
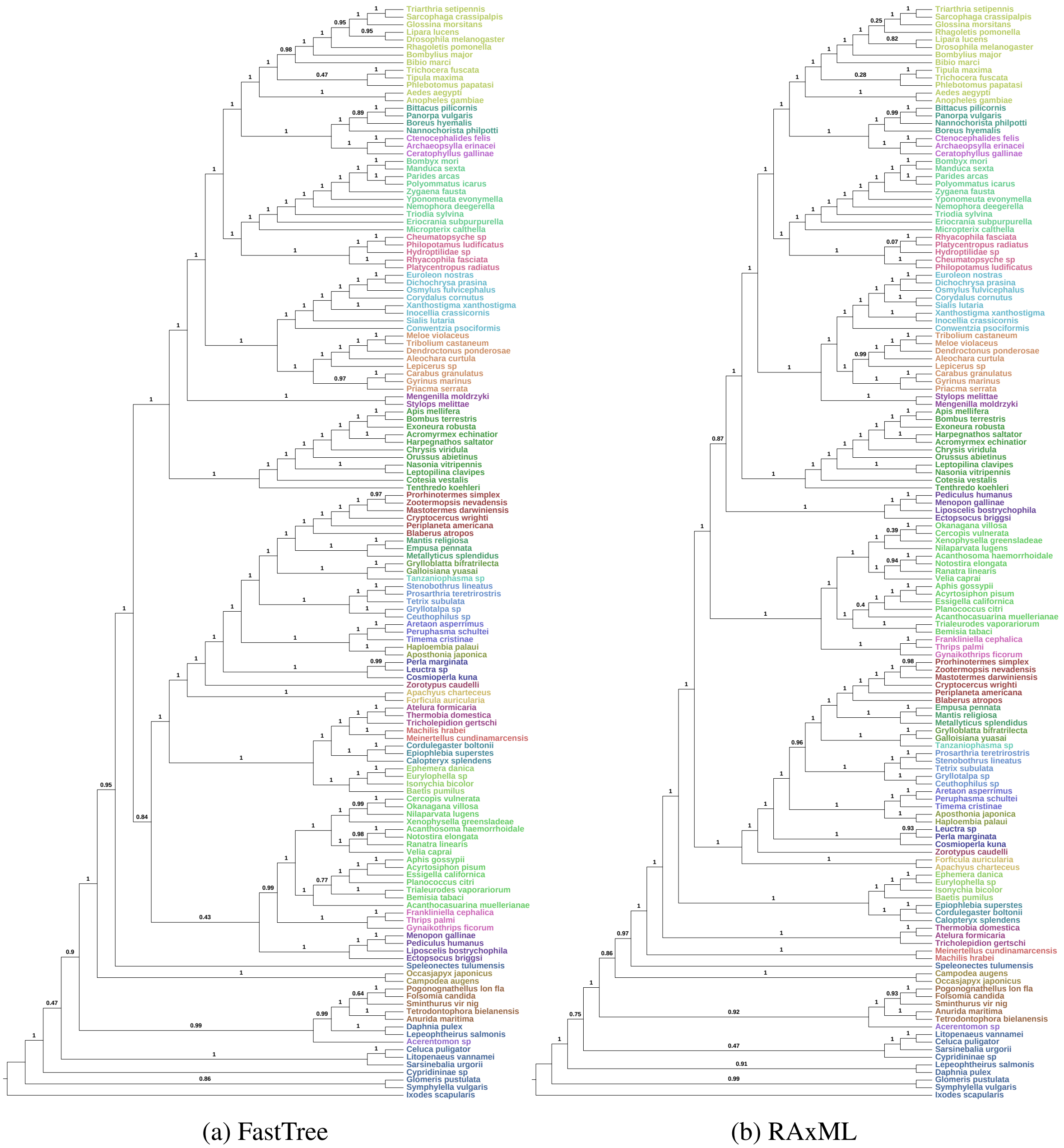


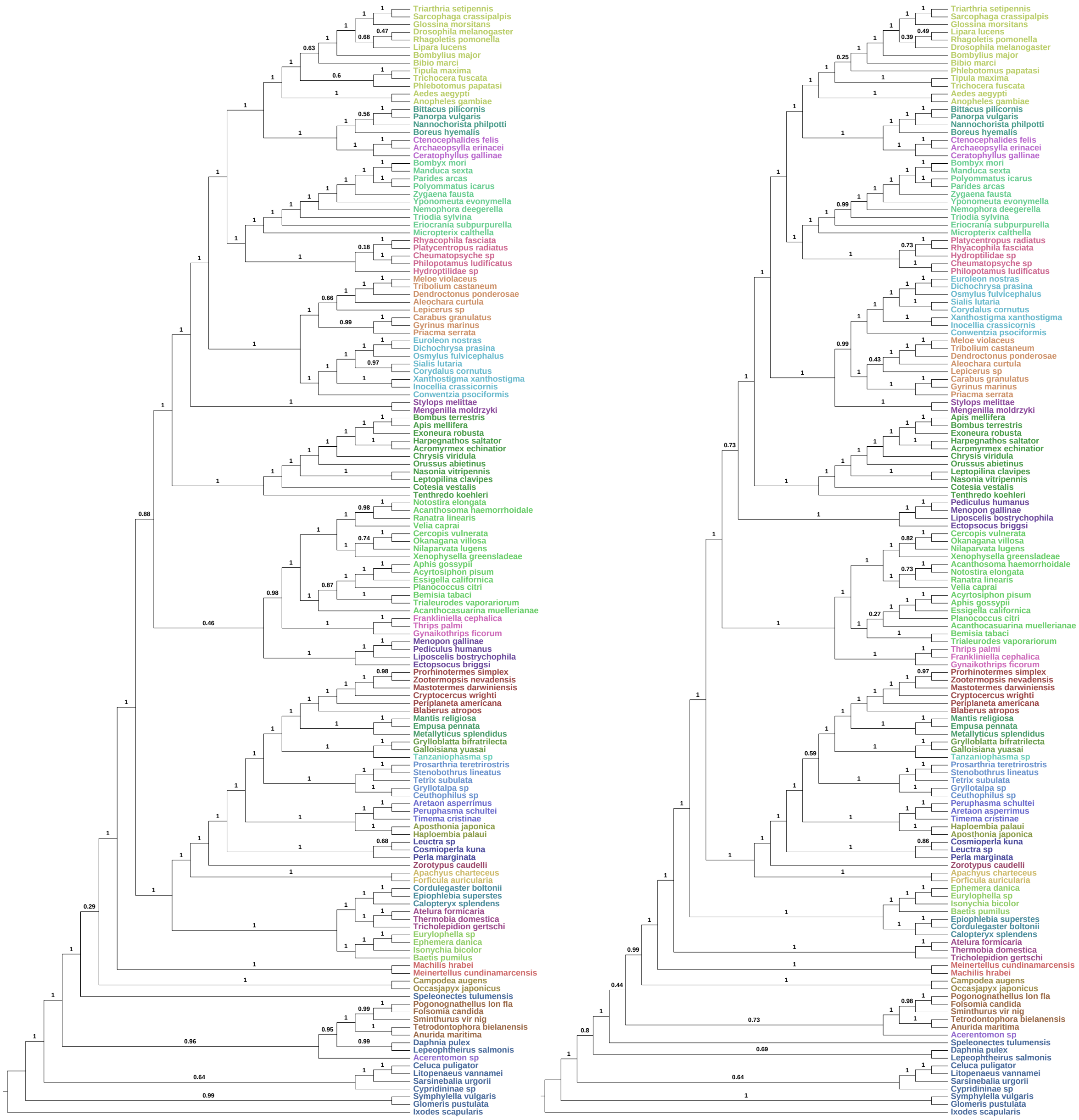
Figure 5.9: Recovery of all clades and important branches in ASTRAL species trees on RAxML and FastTree gene trees with various filtering thresholds (Figs. 5.10 – 5.17, and Fig 5.8), represented by columns. Rows show orders and important clades with letters A to P according to Table 5.2. The blue-green indicates monophyly of a clade, and the spectrum of blue to green colors show ASTRAL localPP support values [210]. Red, indicates strong or weak rejection of a clade. Weak rejection is defined as a clade that is absent from a tree but is compatible with the tree if branches below 95% support are contracted.



(a) FastTree

(b) RAxML

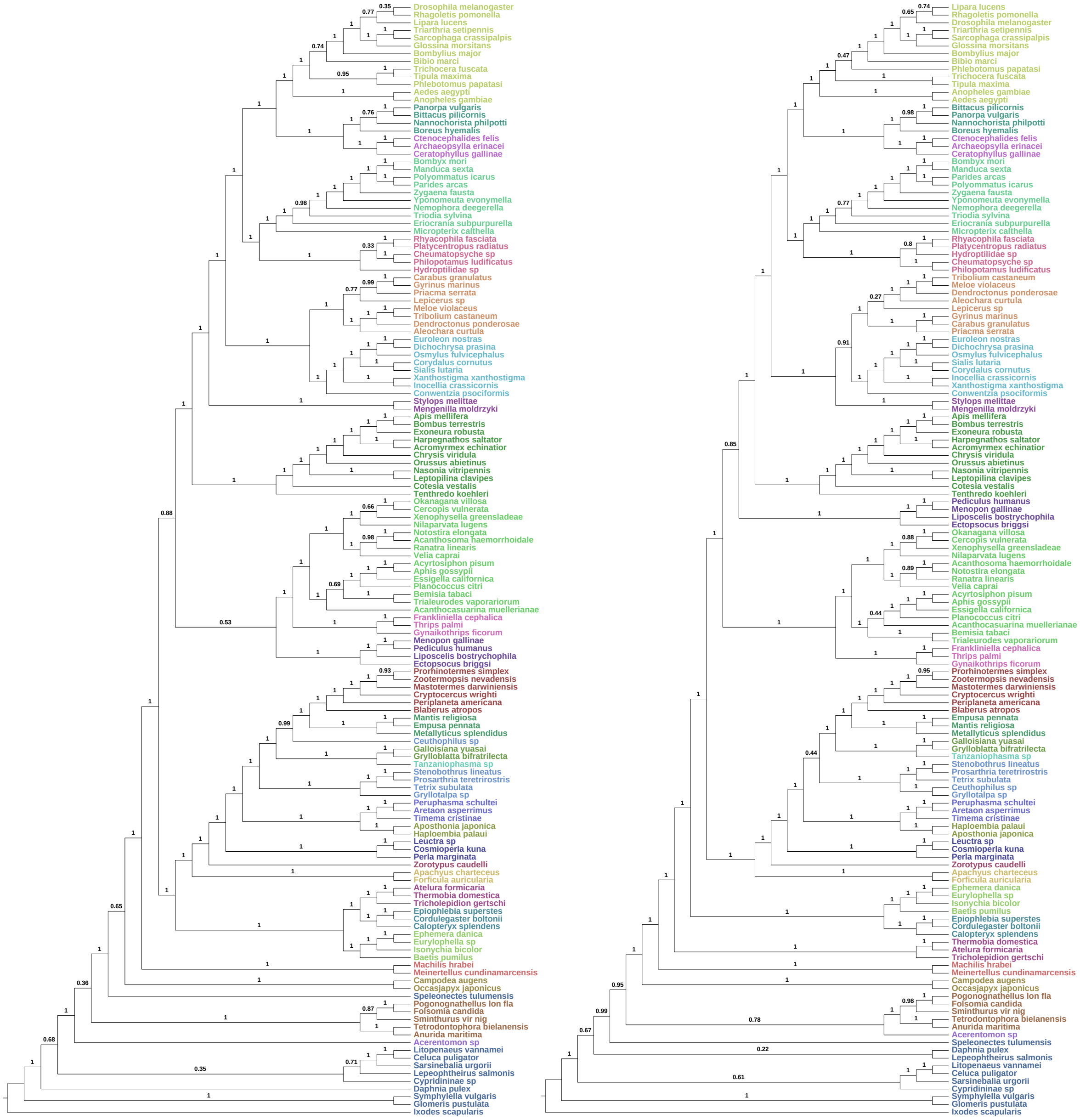
Figure 5.10: ASTRAL species tree on fragmented gene trees without any filtering



(a) FastTree

(b) RAxML

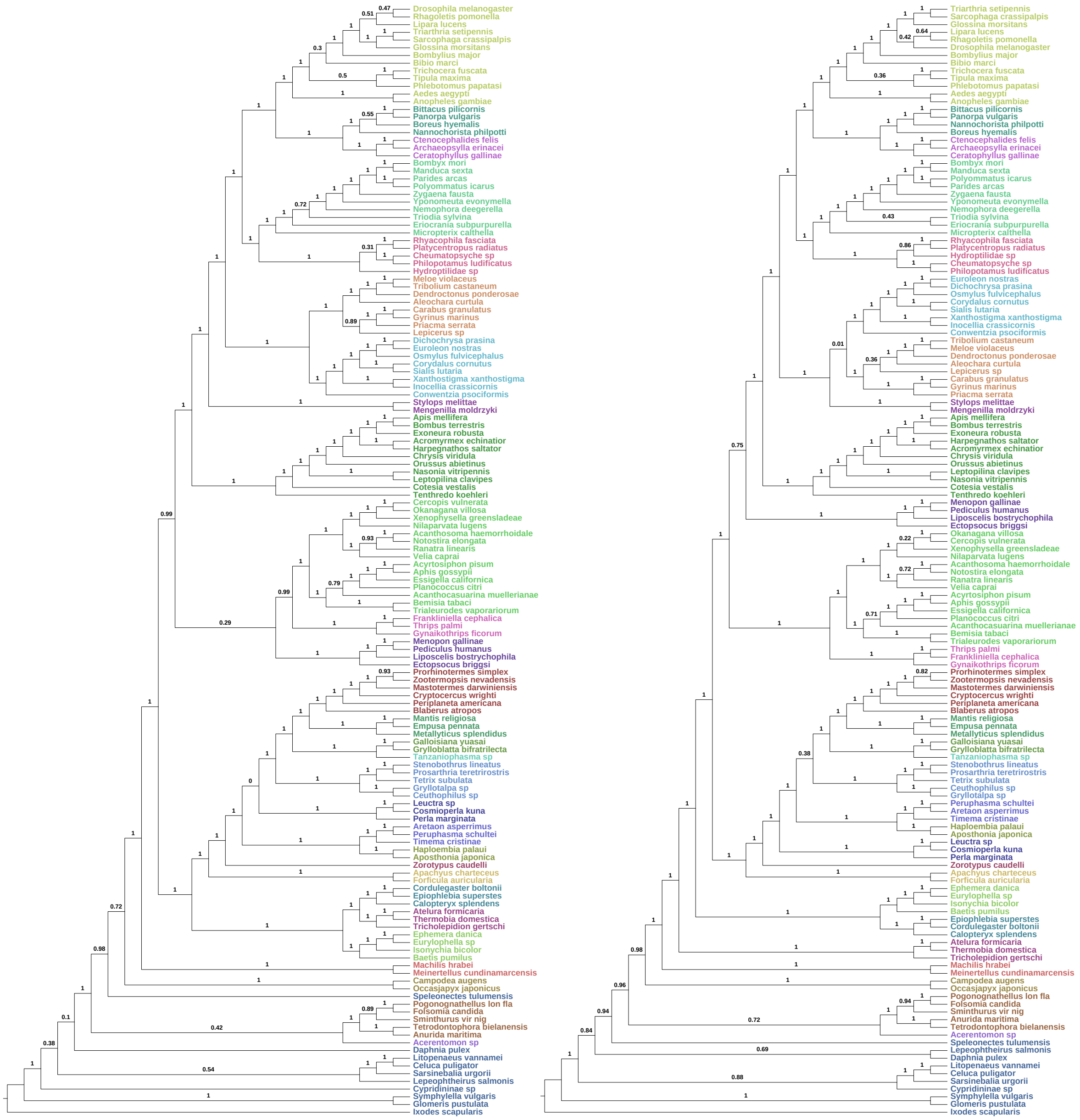
Figure 5.11: ASTRAL species tree with the filtering thresholds of 10% on sites, and 20% on species on 1478 genes.



(a) FastTree

(b) RAxML

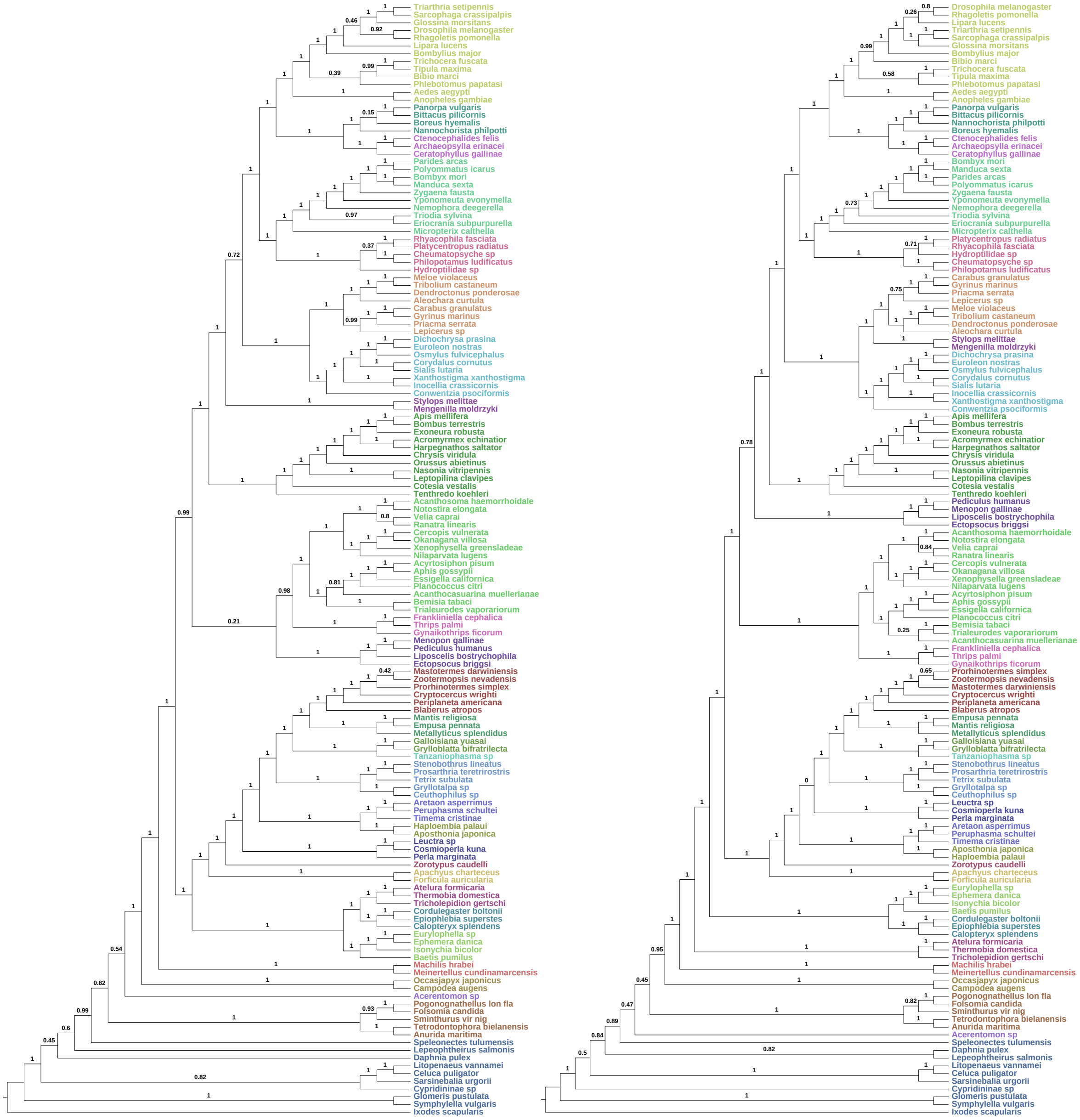
Figure 5.12: ASTRAL species tree with the filtering thresholds of 10% on sites, and 25% on species on 1478 genes.



(a) FastTree

(b) RAxML

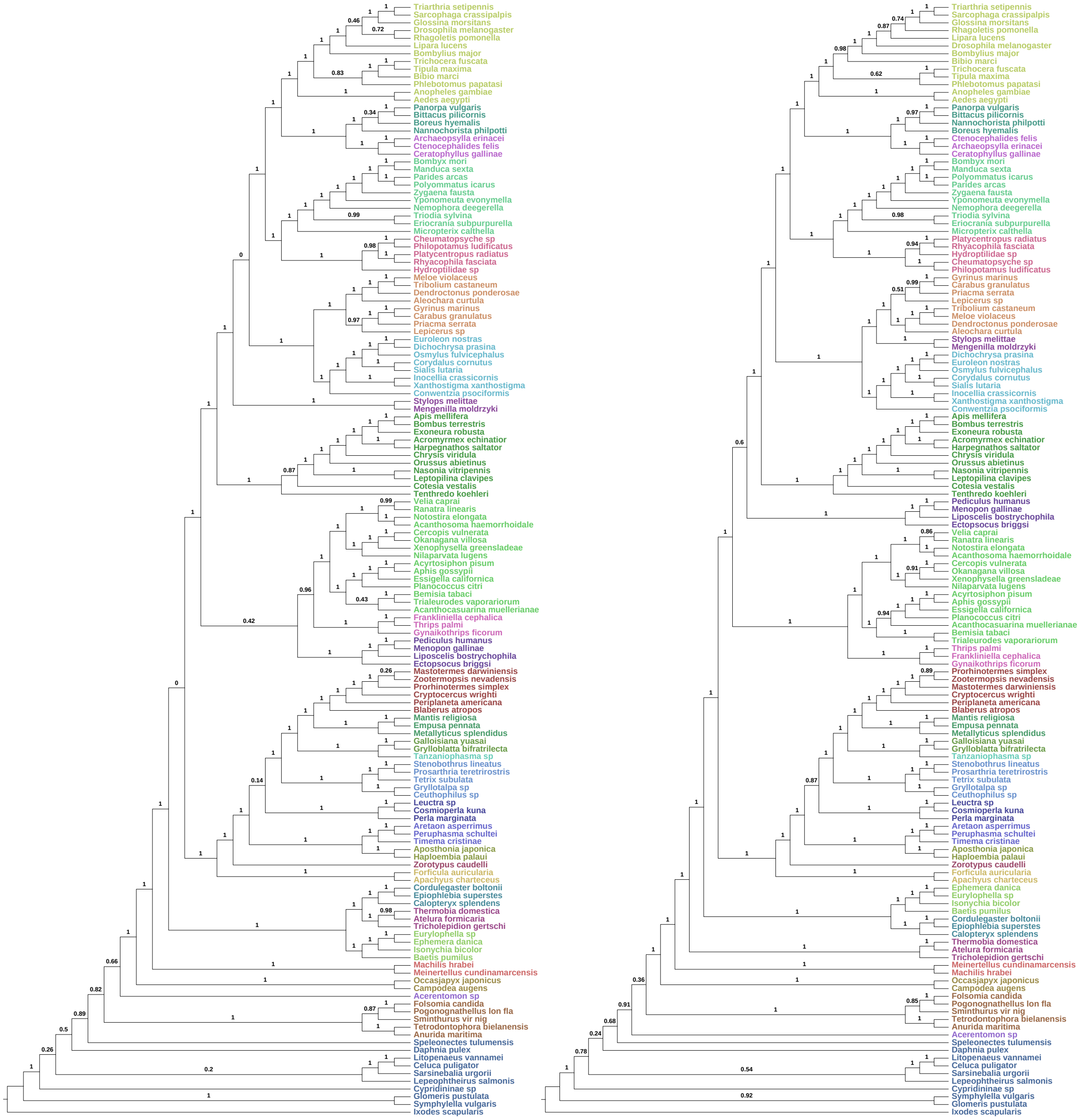
Figure 5.13: ASTRAL species tree with the filtering thresholds of 10% on sites, and 33% on species on 1478 genes.



(a) FastTree

(b) RAxML

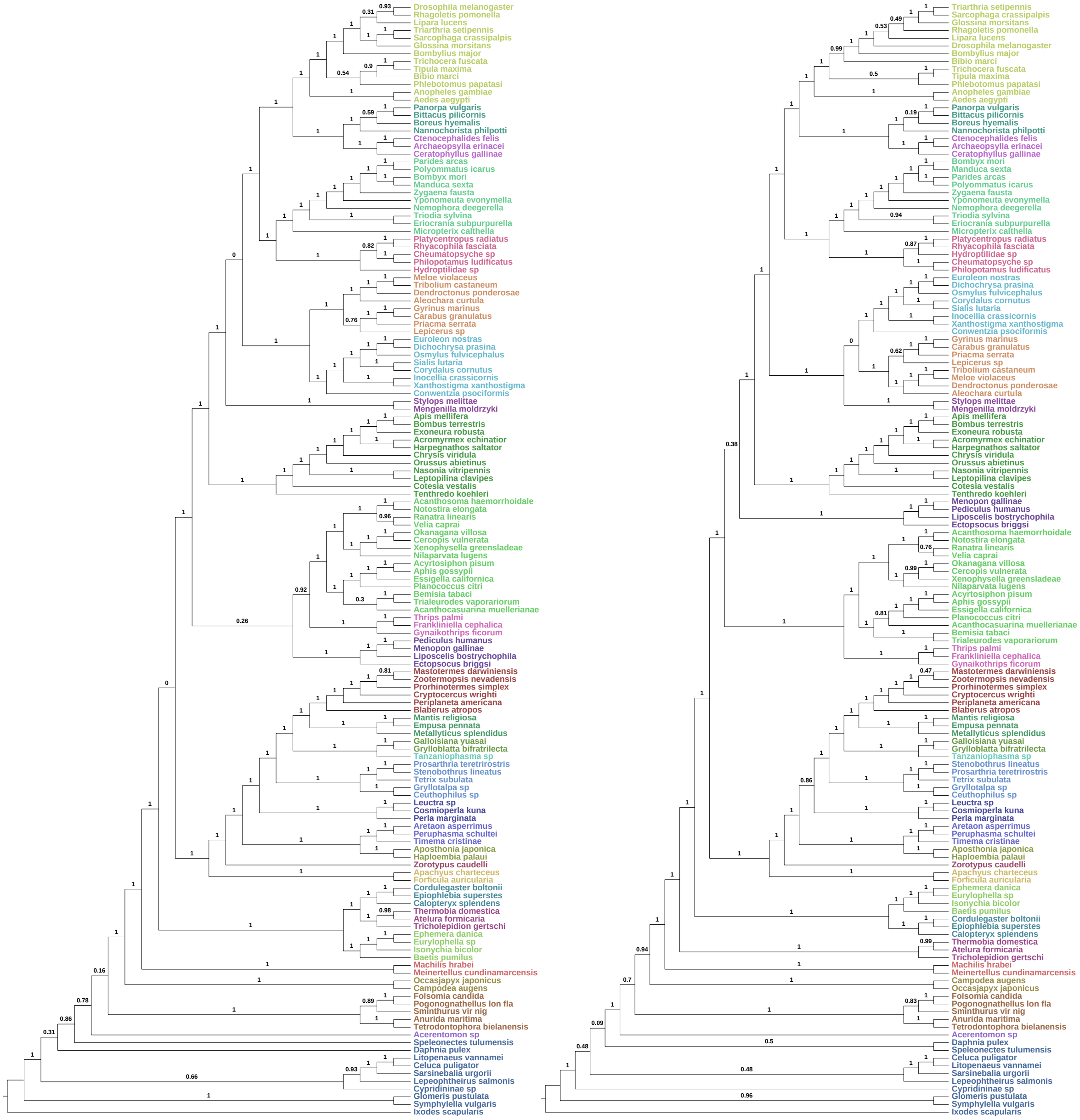
Figure 5.14: ASTRAL species tree with the filtering thresholds of 10% on sites, and 50% on species on 1478 genes.



(a) FastTree

(b) RAxML

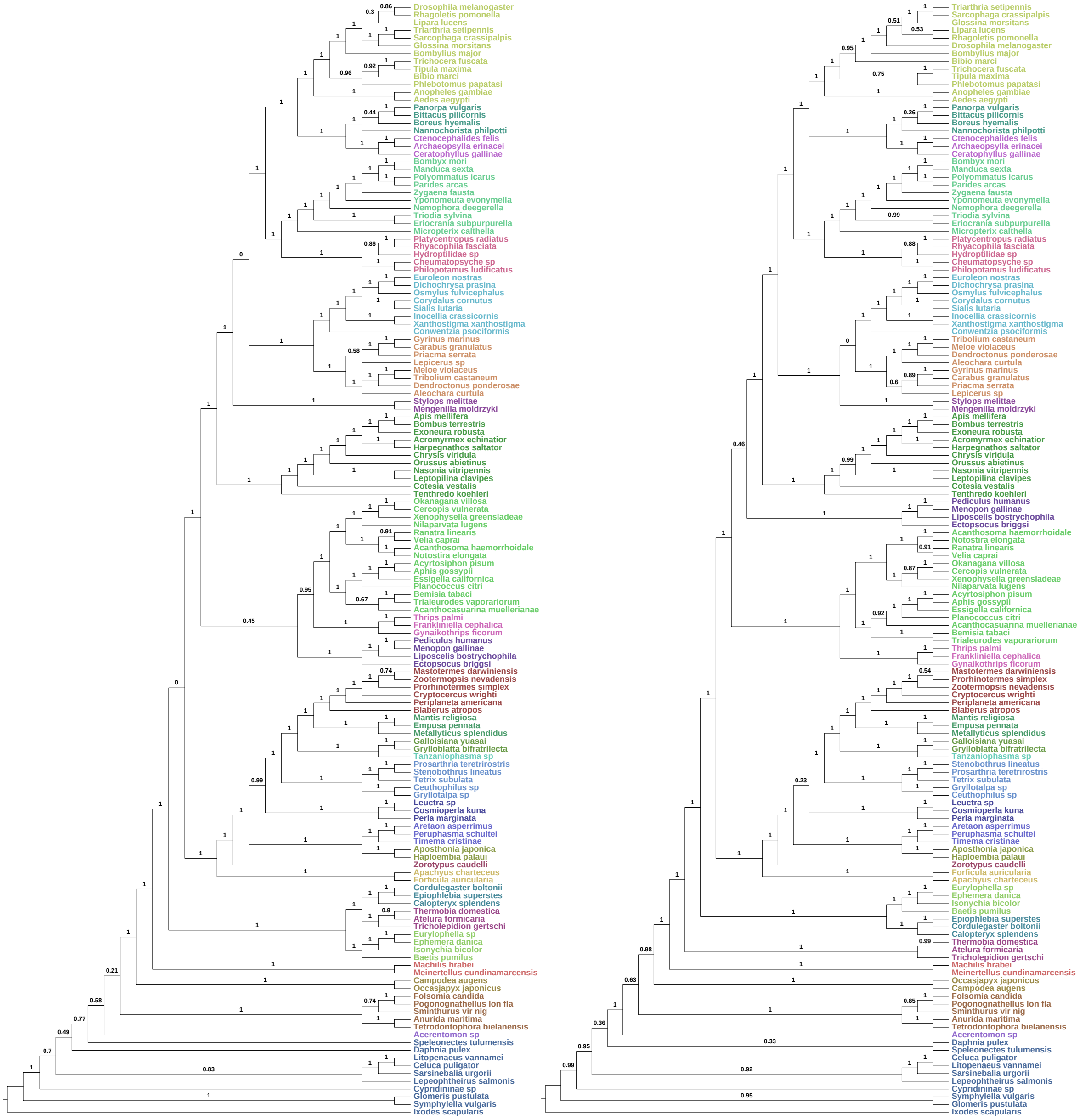
Figure 5.15: ASTRAL species tree with the filtering thresholds of 10% on sites, and 66% on species on 1478 genes.



(a) FastTree

(b) RAxML

Figure 5.16: ASTRAL species tree with the filtering thresholds of 10% on sites, and 75% on species on 1478 genes.



(a) FastTree

(b) RAxML

Figure 5.17: ASTRAL species tree with the filtering thresholds of 10% on sites, and 80% on species on 1478 genes.

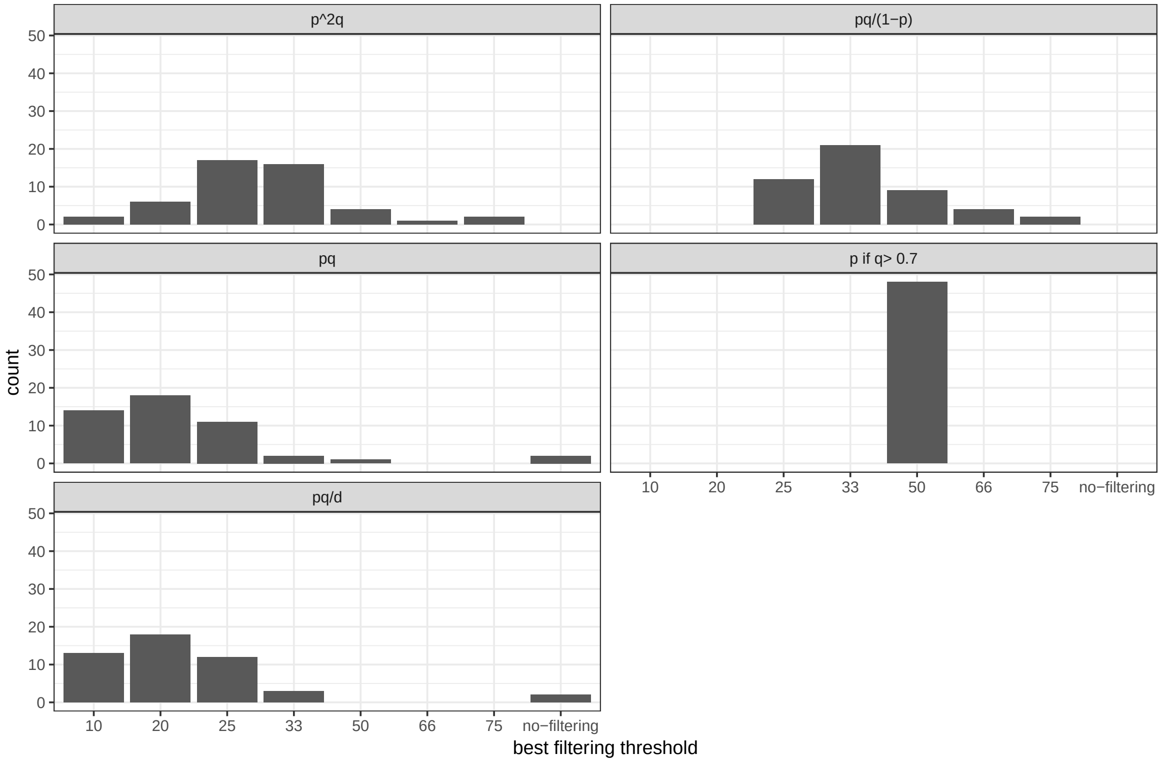


Figure 5.18: Distribution of best filtering threshold according to several rule of thumbs, applied to the simulated dataset with FastTree gene trees. Bootstrapping was not computationally possible for RAxML.

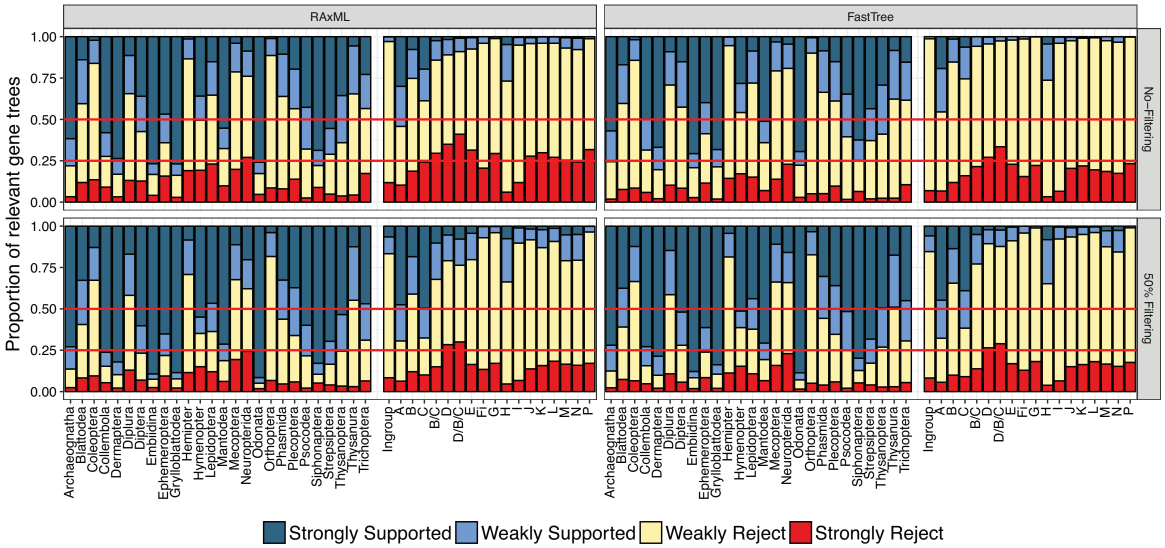


Figure 5.19: Distribution of clade supports for no-filtering and 50%-filtered gene trees. The proportion of (a) RAxML and (b) FastTree gene trees supporting or rejecting each clade is shown. Branches with bootstrap support value 75% and higher (lower) are considered highly (weakly) supported; "Weakly Reject" refers to gene trees that don't recover a clade but are compatible with it once low support branches are contracted; "Strongly Reject" are clades that remain incompatible even after contracting low support branches. Two red horizontal lines mark 50% and 75% support marks.

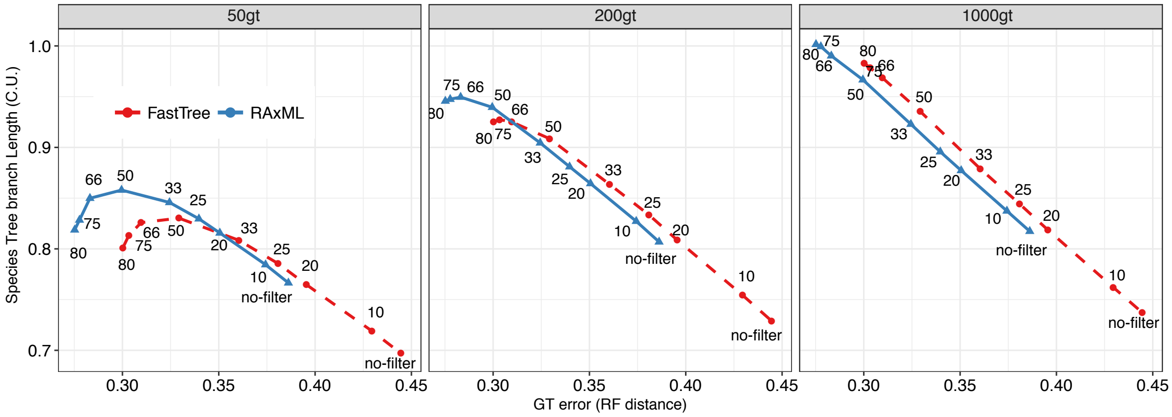


Figure 5.20: Correlation of average ASTRAL species tree branch lengths and average gene tree error (RF distance) when different number of RAxML or FastTree gene trees (50, 200, or 1000) are used. The true average species trees internal branch length is 1.55. Thus, all methods underestimate branch lengths, but filtering reduces the underestimation.

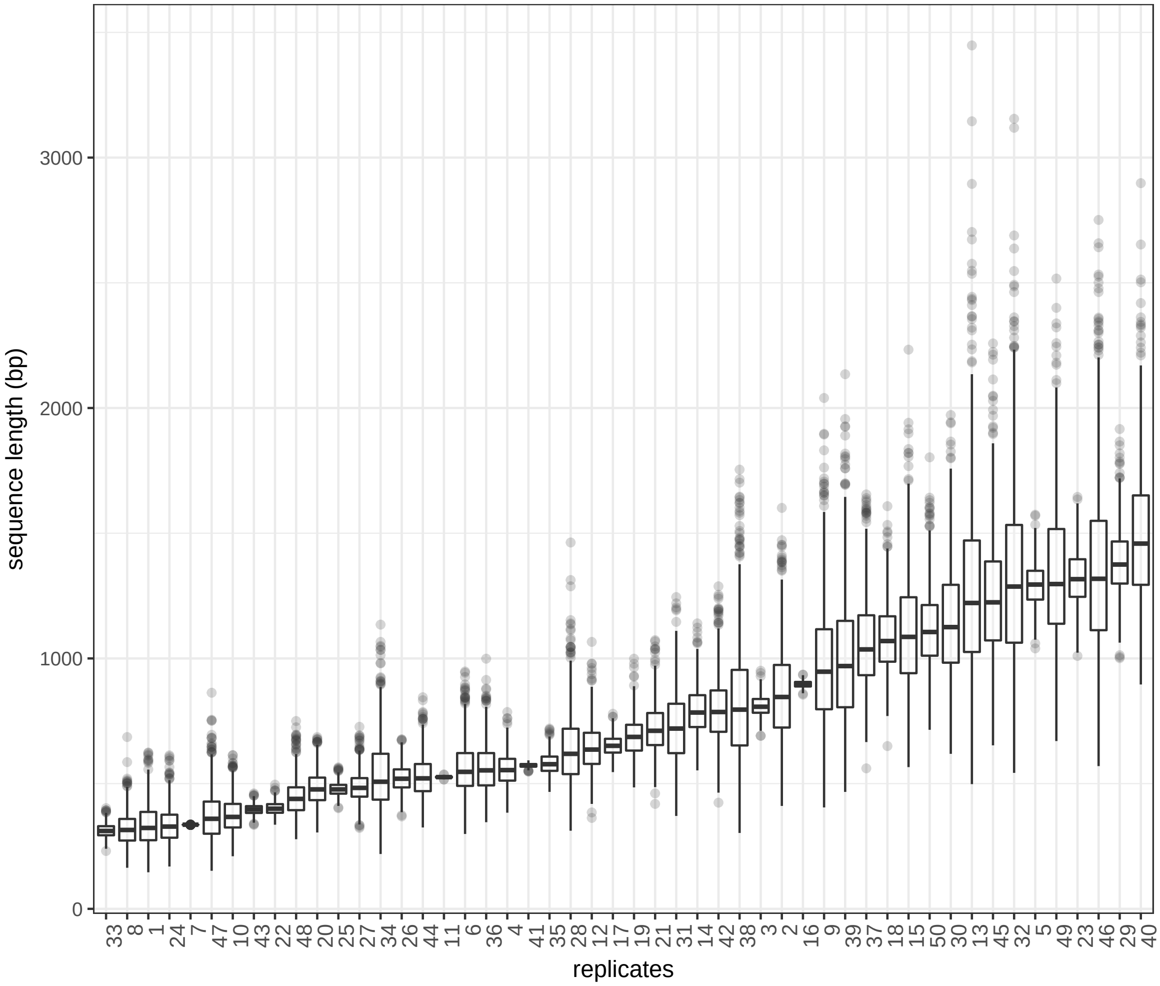


Figure 5.21: The distribution of the original simulated gene alignment lengths, ordered by the median over the replicates.

Unlike FastTree, RAxML implements many protein substitution models and it can find the best scoring protein-coding substitution model [188, 232]. I used RAxML's automatic model selection approach; numbers of genes with various models are shown in Table 5.5. When several species have identical sequences for a gene, I keep only one of them (i.e., remove redundant ones) in my RAxML runs and add the removed species back to the final inferred gene tree as a polytomy.

For performing gene tree bootstrapping using FastTree, I first generate bootstrap sequences using RAxML and then run FastTree on those to estimate the bootstrapped gene trees. I then draw those bootstrap gene trees on ML gene tree branches using the newick utility [110]. For RAxML gene trees, I use the rapid bootstrapping option on reduced sequences (after removing identical sequences). After gene tree estimations, I add back the identical species and draw these bootstrap gene trees on the best ML gene trees (RAxML) following the same procedure using the newick utility.

Species trees: I use ASTRAL-II to estimate the species trees summarizing gene trees with at least 4 taxa left after filtering.

5.4.2 Simulation procedure

I use one model condition of a previously simulated dataset from [165] with 100 ingroup taxa and one outgroup. For each of the 50 replicates in this dataset, Simphy [149] was used to simulate a species tree according to the Yule model, and then 1000 gene trees were simulated using the MSC model which captures ILS. The dataset has moderate levels of ILS; the average distance between true gene trees and true species trees is 0.33. I subsampled genes to create three different datasets with 50, 200, or 1000 genes. DNA sequences of varying length (Fig. 5.21) were simulated down the gene trees using Indelible [72] with GTR parameters and stationary distributions estimated from published biological datasets, as detailed by [165]. Note that simulated sequences did not include any indels and thus were already aligned. [165] suggested removing two replicates that include almost no phylogenetic signal, and I use the same strategy, leaving us with 48 replicates. This creates my unfiltered base dataset.

Adding fragmentation: I add fragmentation to my complete simulated dataset using a procedure that seeks to emulate patterns of fragmentation in the insect biological dataset. 1) For each replicate, I order species in the biological dataset and the simulated dataset with respect to the tip-to-root distances. 2) I randomly select 100 of the biological species and map them to the simulated species with the same position in the order. The main outgroup (*Ixodes scapularis*) in the biological and simulated datasets always map to each other. 3) For each replicate in the simulated dataset, I randomly sample (with replacement) 1000 genes in the insect datasets that have at least 101 species, including the main outgroup. 4) For each species in each simulated gene, I compute the portion of gap sites in the corresponding gene alignment for the corresponding species in the biological data, and remove the same portion of sites in the simulated dataset at random positions. When a species is missing from a gene in the biological dataset, I use the same species from another randomly chosen gene.

Filtering fragments: Although my simulated data do not include indels, injected fragments can create sites that are almost entirely gaps; these sites increase running time but include minimal signal. We, therefore, remove sites with more than 90% gaps, removing between 0.0% and 2.0% (median: 0.1%) of the total number of characters in all sequences. I then remove from each gene any species that has less than a certain fraction (e.g., 10% – 80%) of the full gene. For example, at 10%, I remove only sequences that have 90% or more gaps.

Gene trees and species trees: For each threshold, after filtering, I estimate gene trees using both RAxML [232] version 8.2.9 with two starting trees and FastTree [188] version 2.1.9 Double precision using the GTR+Γ model of sequence evolution [249]. I infer the species tree using ASTRAL-II [165] version 4.11.1, which is a commonly used summary method. I build species trees using all 1000 genes or using randomly chosen subsets of 200 or

50 genes.

5.4.3 Statistical tests

All p-values reported are computed using the Analysis of variance (ANOVA) tests. For impacts on gene trees, I use the gene tree method (RAxML vs FastTree) and filtering thresholds as independent variables. For species tree, to study the impact of presence/absence of fragments, I only include species trees of *orig-seq* and *no-filtering*, and use a binary variable to encode it and use another variable for the number of genes. To study the impact of filtering with sufficiently small thresholds, I restrict the data to those with up to 33% filtering and I use the filtering threshold as a numerical independent variable.

5.5 Acknowledgements

Chapter 5, in full, contains material from SAYYARI, E., WHITFIELD, J. B., AND MIRARAB, S. Fragmentary Gene Sequences Negatively Impact Gene Tree and Species Tree Reconstruction. Molecular Biology and Evolution 34, 12 (2017), 3279–3291. I was the primary investigator and author of this paper.

Chapter 6

DiscoVista: interpretable visualizations of gene tree discordance

Phylogenomics has ushered in an age of discordance. Analyses often reveal abundant discordances among phylogenies of different parts of genomes, as well as incongruences between species trees obtained using different methods or data partitions. Researchers are often left trying to make sense of such incongruences. Interpretive ways of measuring and visualizing discordance are needed, both among alternative species trees and gene trees, especially for specific focal branches of a tree. Here, I introduce DiscoVista, a publicly available tool that creates a suite of simple but interpretable visualizations. DiscoVista helps quantify the amount of discordance and some of its potential causes.

6.1 Introduction

The age of phylogenomics, once hoped to be the end of incongruence in phylogenetic analyses [200], has turned out to be ripe with incongruence [106] and methodological difficulties.

The long-understood theoretical concerns about gene tree incongruence due to incomplete lineage sorting (ILS) [147] have been implicated in many studies (e.g., [105, 264, 241]).

While methods that seek to address gene tree incongruence have been developed, no consensus has emerged as to the choice of the best methodology (e.g., [61, 229]). Nevertheless, phylogenomics studies have to at least consider the possibility of gene tree incongruence and its impacts, a feat made difficult by the noisy estimates of gene trees [229, 180, 160, 210]. Moreover, in some cases, the incongruence itself may be of interest [87]. Even ignoring gene tree discordance, choices of models to apply to the sequences, delineation of data partitions, alignment techniques, or simply software packages used to analyze the data have all proved consequential (e.g., [106, 105, 264, 201, 183, 281]).

These difficulties have compelled some researchers to use several alternative models and methods and then test the sensitivity of results to such choices. Occasionally, analysts choose to also perturb the set of species included, and they often run analyses on different partitions of the data. The analyst hopes for congruence between various analyses that would indicate rubustness of the results to assumptions, but often observes differences. Ideally, the results of *all* analyses should be published, to convey the existence of incongruence in results to the reader.

As long as incongruence remains an important force in phylogenetics, I need interpretable ways to measure and visualize the discordance between species tree estimates resulting from different analytical method and assumptions, and also between gene trees and a summary species tree. Sophisticated tools have been developed to visualize discordance. For example, DensiTree [31] overlays trees on top of each other to create a phylogenetic cloud, and SplitsTree [101] conveys incongruence by producing a network while keeping some of the tree-like structure. These tools create creative and striking indicators of discordance. Yet it is often hard to interpret the meaning of such figures in measurable ways. I believe that in addition to these methods, phylogenomics will benefit from simple, interpretable, and

Table 6.1: Description and examples for different DiscoVista analyses

Analysis Name	**Shows . . .**	**Examples**
Species tree compatibility	compatibility of focal groups of species in species tree	Figs. 6.8a, 6.1, and 6.2
Occupancy analysis	taxon occupancy for individuals focal groups of species in genes	Figs. 6.8bc and 6.3
Gene tree compatibility	compatibility of focal groups of species in gene trees	Figs. 6.9a, 6.10, and 6.11
Branch quartet frequencies	quartet frequencies around important branches of the species tree	Figs. 6.9b and 6.4
GC content	GC content of each codon position	Figs. 6.5

easy-to-perform visualizations that help systematists to identify discordance and its potential causes.

in this dissertation, I introduce DiscoVista, a tool that creates a series of simple yet powerful visualizations of discordance. DiscoVista is a command-line tool and relies on several other packages, including Dendropy [242], ape [179], newick utilities [110], the ggplot package [265], and ASTRAL [210]. The code, a Docker [29] virtual image (for easy installation), and examples are available online at `https://github.com/esayyari/DiscoVista`.

DiscoVista can generate several visualizations (Table 6.1) that summarize gene tree discordance and discordance among species trees, show taxon occupancy, and show sequence statistics such as GC content. DiscoVista strives for interpretability. In many analyses, not all branches in a phylogenetic tree are equally important because questions of interest typically concern several hypotheses surrounding the relationships between focal groups. Visualizing discordance with respect to only these focal relationships simplifies interpretation. Assessing hypotheses concerning these larger subsets of the species helps in answering the downstream biological questions of interest. Thus, DiscoVista allows researchers to define focal groups of taxa and evaluate discordance relevant to those groups.

6.2 Results

I apply DiscoVista on three datasets to demonstrate its output visualizations. The exact commands for generating example figures are given in the appendixs.

Datasets The position of Xenacoelomorpha among deep branches of the Metazoan phylogeny has proved challenging to resolve, with two prevailing hypotheses. One hypothesis puts Xenacoelomorpha as sister to all other Bilateria, while the other hypothesis puts them inside Deuterostomia, implying a dramatic loss of complexity. Intriguingly, these marine worms are bilaterally symmetrical but lack several other features compared to most other bilaterians. Two independent and simultaneous studies by [203] and [43] have focused on the position of Xenacoelomorpha in the tree of life. These two studies used different (but overlapping) set of species and each analysis used several reconstruction methods. The two papers come to the same conclusion, putting Xenacoelomorpha as the sister to all other Bilateria. The final results of [43] is based on 78 species and 212 orthologous genes with average per taxon occupancy of 80%. Their paper includes alternative analyses based on several subsets of taxa and reconstruction methods. The dataset by [203] includes 26 species and 1178 genes (including four Xenoturbella species) with average gene occupancy of 70%. They also report alternative trees using concatenation and ASTRAL (run on 393 genes with 80% occupancy). Although these two datasets used different set of taxa, by focusing on focal splits, DiscoVista can generate visual comparisons of results across both datasets (Figs. 6.8 and 6.9).

As a second example, I show DiscoVista results on a phylotranscriptomic dataset of 103 plants [264] in the appendix (Figs. 6.1 – 6.5). This dataset comes with both DNA and AA sequences, allowing us to show additional figures that could not be built for the Xenacoelomorpha datasets.

Split definitions A central input to DiscoVista is a *split definitions* file where the user

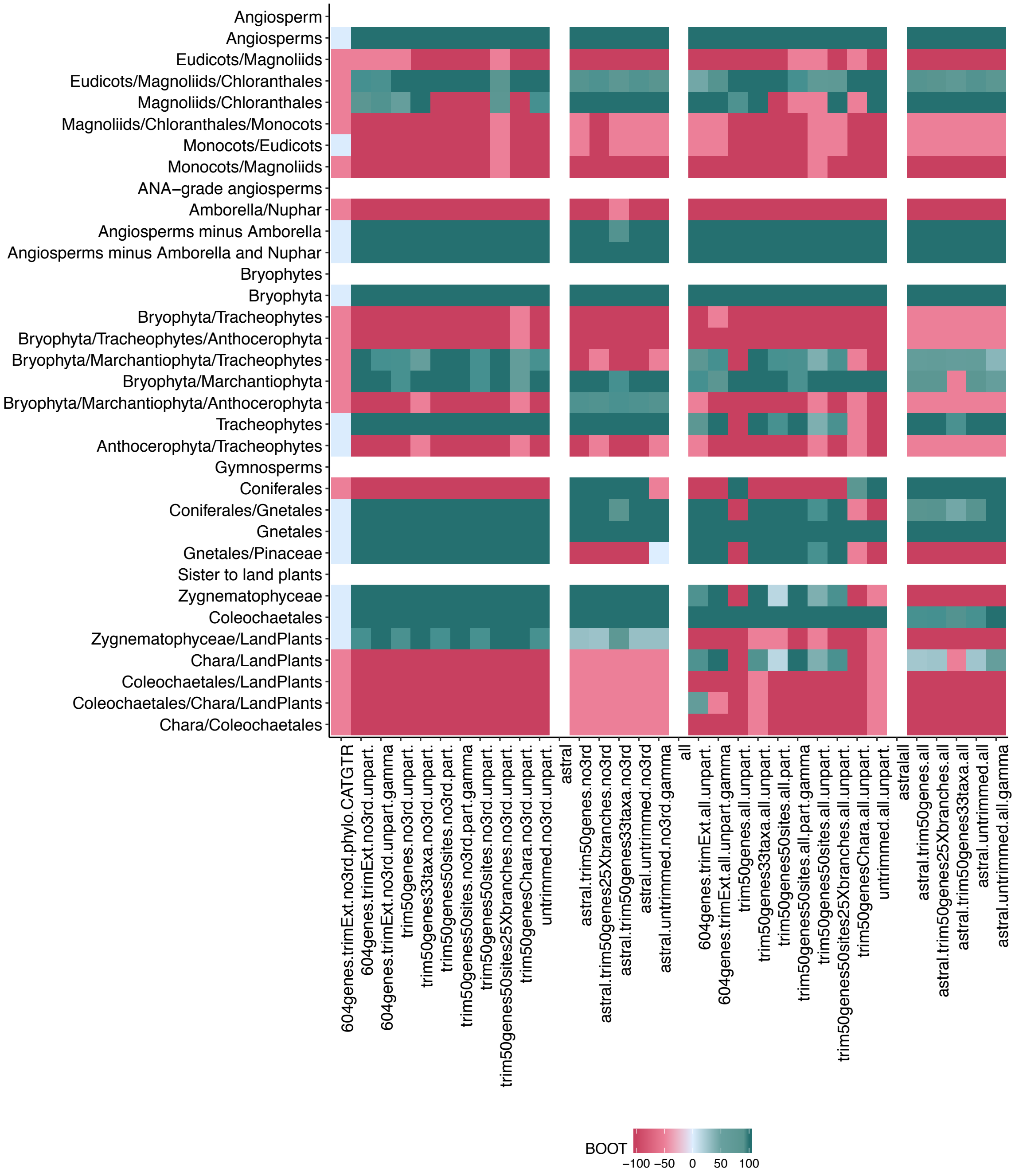


Figure 6.1: DiscoVista Specie tree analysis on 1kp dataset: Rows correspond to major orders and clades, and columns correspond to the results of different methods reported in two different closely related datasets. The spectrum of blue-green indicates amount of MLBS values for monophyletic clades. Weakly rejected clades correspond to clades that are not present in the tree, but are compatible if low support branches (below 90%) are contracted.

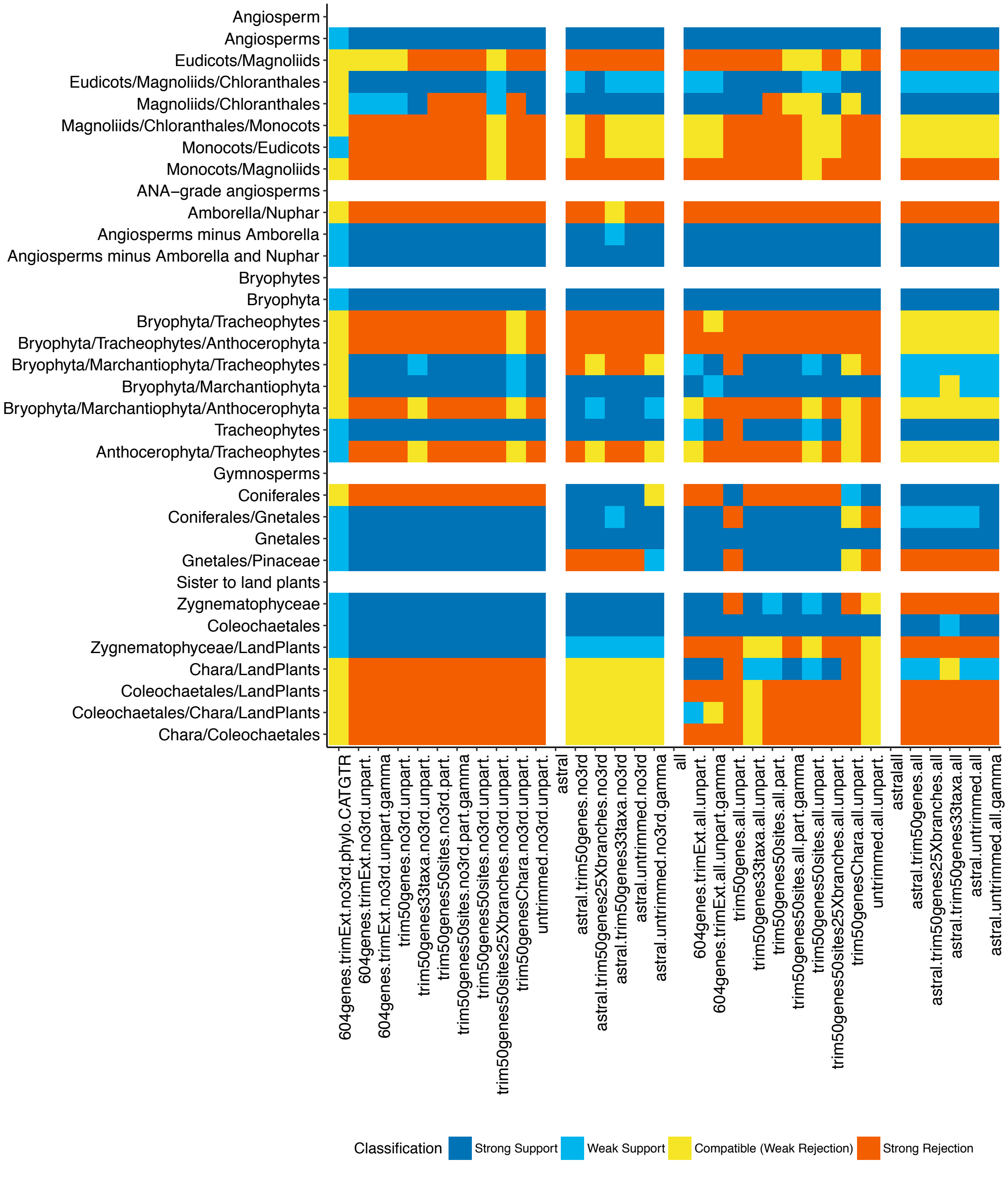


Figure 6.2: DiscoVista Specie tree analysis on 1kp dataset: Rows correspond to major orders and clades, and columns correspond to the results of different methods reported in two different closely related datasets. Weakly rejected clades correspond to clades that are not present in the tree, but are compatible if low support branches (below 90%) are contracted.

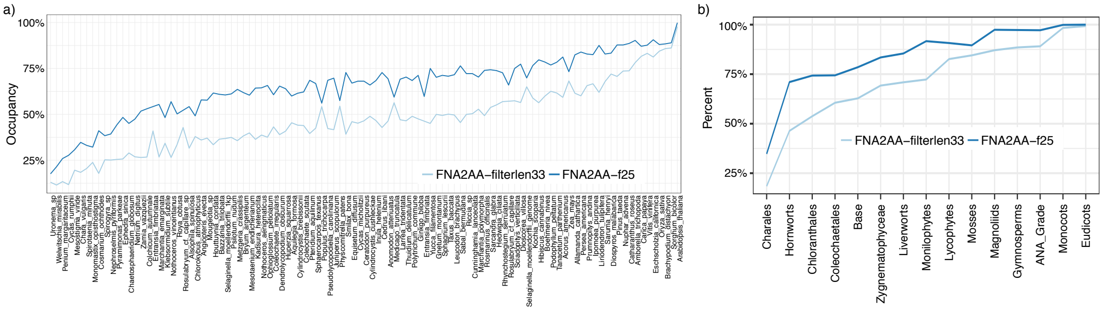


Figure 6.3: DiscoVista analyses on 1kp dataset a) occupancy analysis on the 1kp dataset over each individual species for two model conditions. FNA2AA-f25: amino acid sequences back translated to DNA, and sequences on long branches (25X median branch length) removed; FNA2AA-filterlen33: amino acids sequences back translated to DNA, and fragmentary sequences removed (66% gaps or more). b) occupancy analysis of major clades in the 1kp dataset and the same model conditions as (a).

can combine taxa into groups of interest and give names to the groups (see supplementary material for details). Each split is a bipartition of the taxa into two groups and corresponds to an edge in an unrooted tree. The user can specify one side of a split (which would be a clade if the side that doesn't include the root is given). With careful definition of splits, alternative hypotheses of interest could be specified. For my two empirical datasets, I am considering focal groups of species from original publications [203, 43], as shown partially in Table 6.2 (see Figs. 6.6 and 6.7 for full definitions).

i) Species tree compatibility This visualization shows whether focal splits are supported, weakly rejected, or strongly rejected by different analyses. The inputs are a set of species trees in newick format, a support threshold, and the splits definition file; the output is a heat map. For each focal split and for each species tree, if the split is compatible with the tree, the corresponding cell is in shades of blue/green, and the spectrum of blue-green color indicates the branch support (any measure of support, including the bootstrap support, Bayesian posterior probability, or localPP [210] values could be used.) A split is considered weakly rejected if it is incompatible with the fully resolved species tree, but is *compatible*

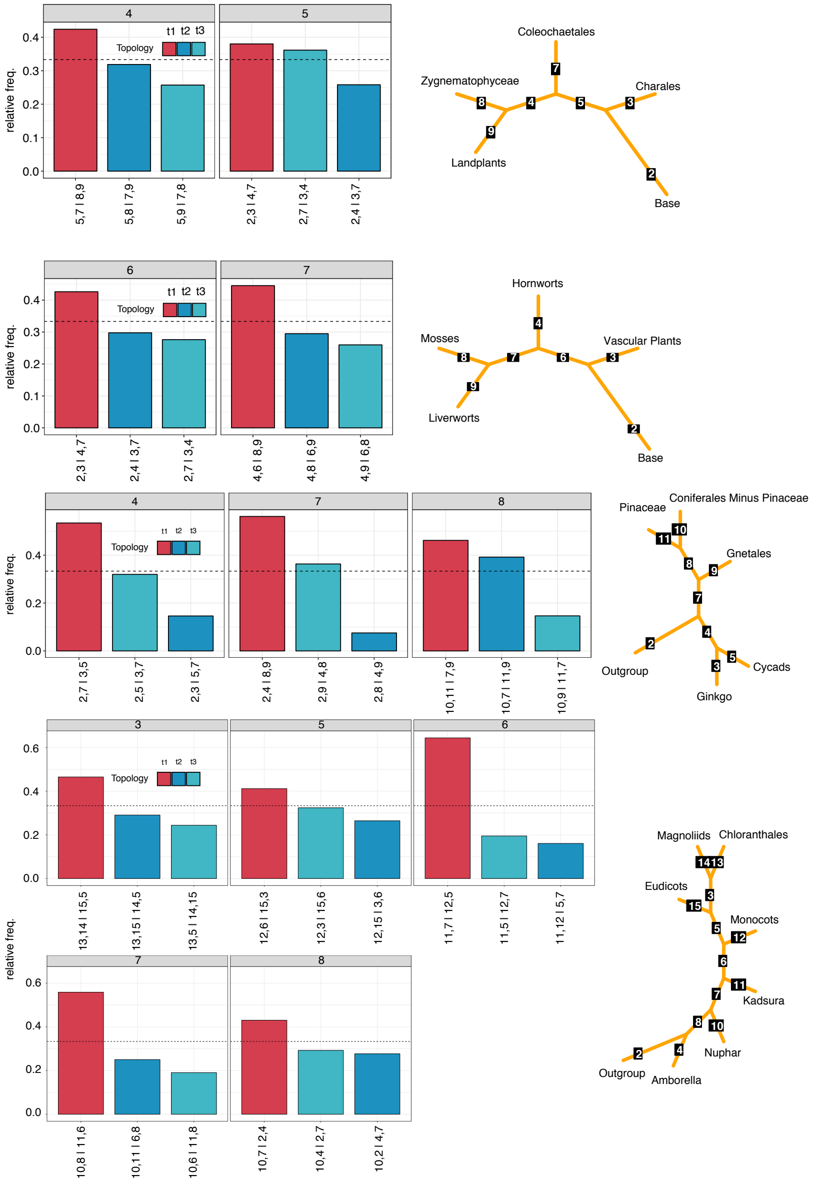


Figure 6.4: Example DiscoVista visualizations of gene tree discordance on the 1kp dataset [264]. Branch quartet frequencies graphs. Bars show the frequencies of observing the three quartet topologies around focal branches of the ASTRAL species tree among the main trimmed gene trees of the 1kp dataset. Each internal branch has four neighboring branches, which can be arranged in three ways. The frequency of the species tree topology among gene trees is shown in red, and the other two alternative topologies are shown in blue. The dotted lines indicate the 1/3 threshold. The title of each box indicates the label of the corresponding branch on the associated cartoon tree (also generated by DiscoVista). On the x-axis the exact definition of each quartet topology is shown using the neighboring branch labels separated by “#”.

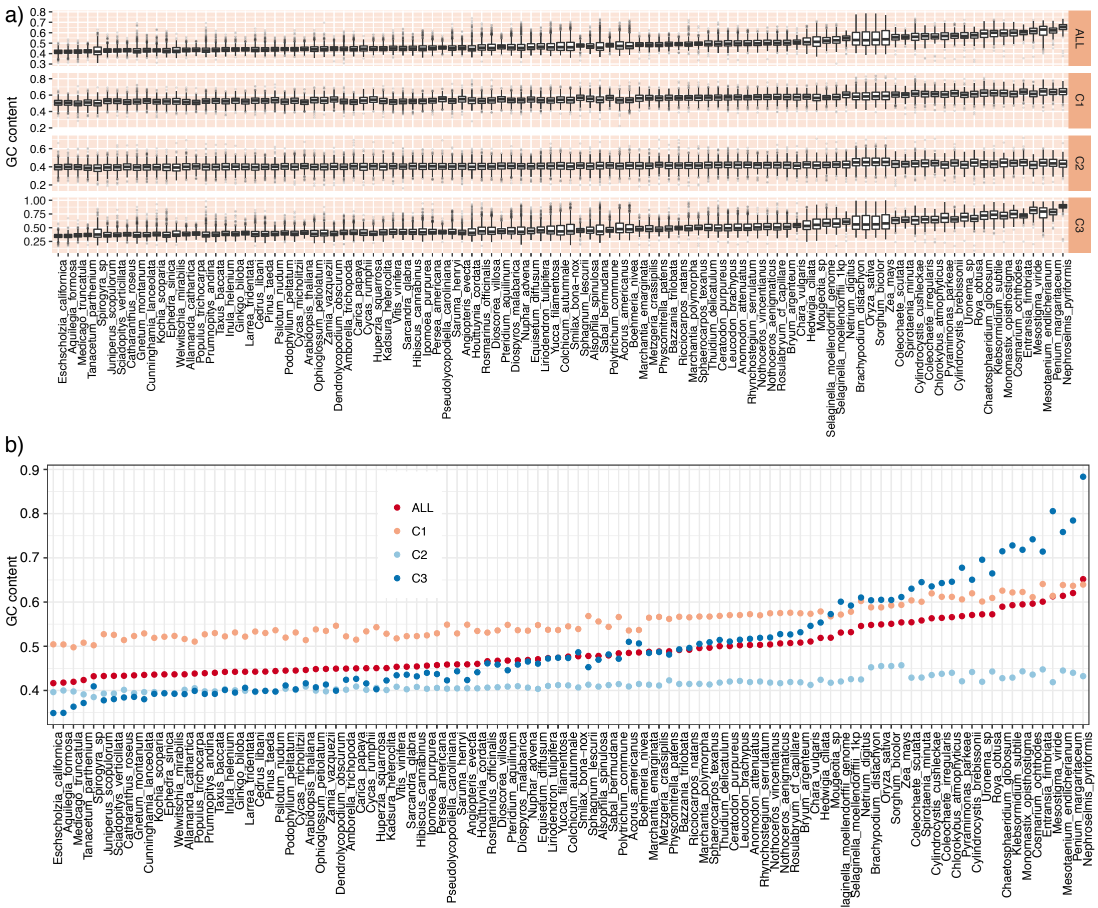
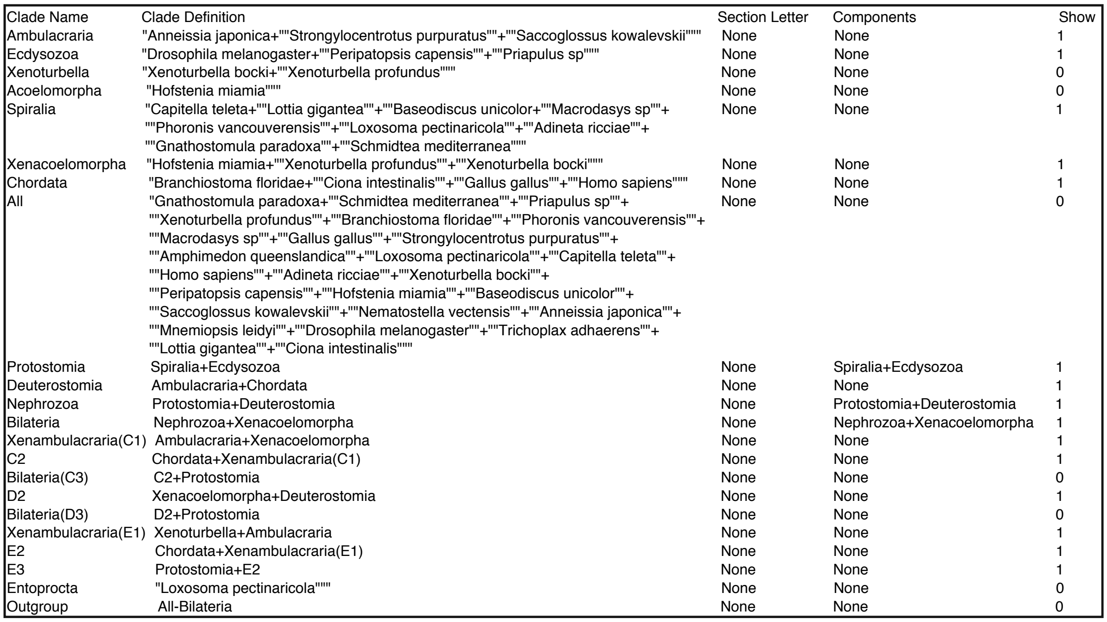


Figure 6.5: DiscoVista analyses on 1kp dataset a,b) GC content analysis of the 1kp dataset using boxplots and dot plot respectively for first, second, third, as well as all three codon positions. In dot-plot each dot shows the average GC content ratio for each species in all (red), first (pink), second (light blue), and third (dark blue) codon positions.

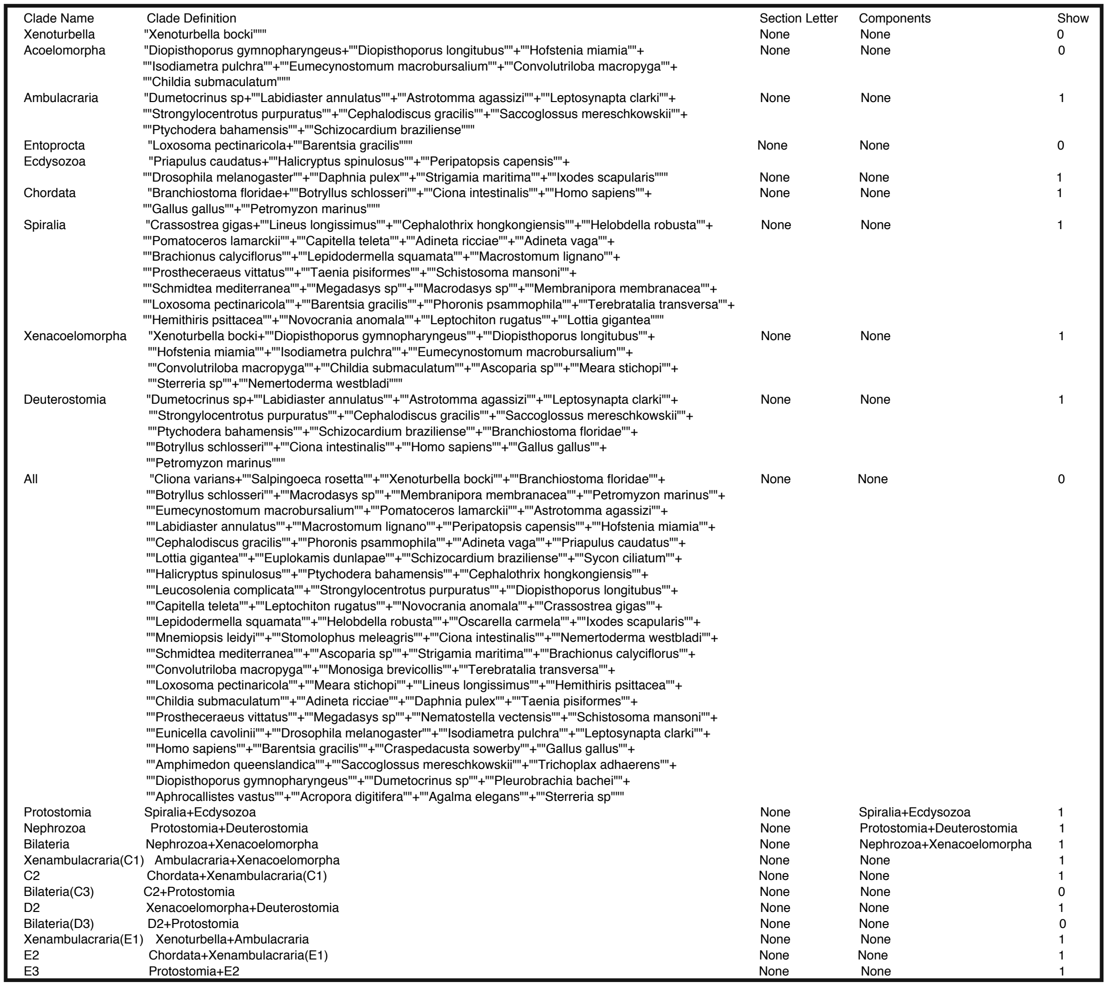


Clade Name	Clade Definition	Section Letter	Components	Show
Ambulacraria	"Anneissia japonica+""Strongylocentrotus purpuratus""+""Saccoglossus kowalevskii"""	None	None	1
Ecdysozoa	"Drosophila melanogaster+""Peripatopsis capensis""+""Priapulus sp"""	None	None	1
Xenoturbella	"Xenoturbella bocki+""Xenoturbella profundus"""	None	None	0
Acoelomorpha	"Hofstenia miamia"""	None	None	0
Spiralia	"Capitella teleta+""Lottia gigantea""+""Baseodiscus unicolor+""Macrodasys sp""+ ""Phoronis vancouverensis""+""Loxosoma pectinaricola""+""Adineta ricciae""+ ""Gnathostomula paradoxa""+""Schmidtea mediterranea"""	None	None	1
Xenacoelomorpha	"Hofstenia miamia+""Xenoturbella profundus""+""Xenoturbella bocki"""	None	None	1
Chordata	"Branchiostoma floridae+""Ciona intestinalis""+""Gallus gallus""+""Homo sapiens"""	None	None	1
All	"Gnathostomula paradoxa+""Schmidtea mediterranea""+""Priapulus sp""+ ""Xenoturbella profundus""+""Branchiostoma floridae""+""Phoronis vancouverensis""+ ""Macrodasys sp""+""Gallus gallus""+""Strongylocentrotus purpuratus""+ ""Amphimedon queenslandica""+""Loxosoma pectinaricola""+""Capitella teleta""+ ""Homo sapiens""+""Adineta ricciae""+""Xenoturbella bocki""+ ""Peripatopsis capensis""+""Hofstenia miamia""+""Baseodiscus unicolor""+ ""Saccoglossus kowalevskii""+""Nematostella vectensis""+""Anneissia japonica""+ ""Mnemiopsis leidyi""+""Drosophila melanogaster""+""Trichoplax adhaerens""+ ""Lottia gigantea""+""Ciona intestinalis"""	None	None	0
Protostomia	Spiralia+Ecdysozoa	None	Spiralia+Ecdysozoa	1
Deuterostomia	Ambulacraria+Chordata	None	None	1
Nephrozoa	Protostomia+Deuterostomia	None	Protostomia+Deuterostomia	1
Bilateria	Nephrozoa+Xenacoelomorpha	None	Nephrozoa+Xenacoelomorpha	1
Xenambulacraria(C1)	Ambulacraria+Xenacoelomorpha	None	None	1
C2	Chordata+Xenambulacraria(C1)	None	None	1
Bilateria(C3)	C2+Protostomia	None	None	0
D2	Xenacoelomorpha+Deuterostomia	None	None	1
Bilateria(D3)	D2+Protostomia	None	None	0
Xenambulacraria(E1)	Xenoturbella+Ambulacraria	None	None	1
E2	Chordata+Xenambulacraria(E1)	None	None	1
E3	Protostomia+E2	None	None	1
Entoprocta	"Loxosoma pectinaricola"""	None	None	0
Outgroup	All-Bilateria	None	None	0

Figure 6.6: Complete split definitions for the Xenoturbella dataset of [203].

Table 6.2: An example *split definition* file for the two Xenoturbella datasets; several lines on top define base splits and are left blank here, but see the full files in Figures 6.6 and 6.7.

Cluster Name	Definition
Ambulacraria	...
Ecdysozoa	...
Acoelomorpha	...
Spiralia	...
Xenacoelomorpha	...
Chordata	...
Protostomia	Spiralia+Ecdysozoa
Deuterostomia	Ambulacraria+Chordata
Nephrozoa	Protostomia+Deuterostomia
Bilateria	Nephrozoa+Xenacoelomorpha
Xenambulacraria(C1)	Ambulacraria+Xenacoelomorpha
C2	Chordata+Xenambulacraria(C1)
Bilateria(C3)	C2+Protostomia
D2	Xenacoelomorpha+Deuterostomia
Bilateria(D3)	D2+Protostomia
Xenambulacraria(E1)	Xenoturbella+Ambulacraria
E2	Chordata+Xenambulacraria(E1)
E3	Protostomia+E2



Clade Name	Clade Definition	Section Letter	Components	Show
Xenoturbella	"Xenoturbella bocki"""	None	None	0
Acoelomorpha	"Diopisthoporus gymnopharyngeus+""Diopisthoporus longitubus""+""Hofstenia miamia""+ ""Isodiametra pulchra""+""Eumecynostomum macrobursalium""+""Convolutriloba macropyga""+ ""Childia submaculatum"""	None	None	0
Ambulacraria	"Dumetocrinus sp+""Labidiaster annulatus""+""Astrotomma agassizi""+""Leptosynapta clarki""+ ""Strongylocentrotus purpuratus""+""Cephalodiscus gracilis""+""Saccoglossus mereschkowskii""+ ""Ptychodera bahamensis""+""Schizocardium braziliense"""	None	None	1
Entoprocta	"Loxosoma pectinaricola+""Barentsia gracilis"""	None	None	0
Ecdysozoa	"Priapulus caudatus+""Halicryptus spinulosus""+""Peripatopsis capensis""+ ""Drosophila melanogaster""+""Daphnia pulex""+""Strigamia maritima""+""Ixodes scapularis"""	None	None	1
Chordata	"Branchiostoma floridae+""Botryllus schlosseri""+""Ciona intestinalis""+""Homo sapiens""+ ""Gallus gallus""+""Petromyzon marinus"""	None	None	1
Spiralia	"Crassostrea gigas+""Lineus longissimus""+""Cephalothrix hongkongiensis""+""Helobdella robusta""+ ""Pomatoceros lamarckii""+""Capitella teleta""+""Adineta ricciae""+""Adineta vaga""+ ""Brachionus calyciflorus""+""Lepidodermella squamata""+""Macrostomum lignano""+ ""Prostheceraeus vittatus""+""Taenia pisiformes""+""Schistosoma mansoni""+ ""Schmidtea mediterranea""+""Megadasys sp""+""Macrodasys sp""+""Membranipora membranacea""+ ""Loxosoma pectinaricola""+""Barentsia gracilis""+""Phoronis psammophila""+""Terebratalia transversa""+ ""Hemithiris psittacea""+""Novocrania anomala""+""Leptochiton rugatus""+""Lottia gigantea"""	None	None	1
Xenacoelomorpha	"Xenoturbella bocki+""Diopisthoporus gymnopharyngeus""+""Diopisthoporus longitubus""+ ""Hofstenia miamia""+""Isodiametra pulchra""+""Eumecynostomum macrobursalium""+ ""Convolutriloba macropyga""+""Childia submaculatum""+""Ascoparia sp""+""Meara stichopi""+ ""Sterreria sp""+""Nemertoderma westbladi"""	None	None	1
Deuterostomia	"Dumetocrinus sp+""Labidiaster annulatus""+""Astrotomma agassizi""+""Leptosynapta clarki""+ ""Strongylocentrotus purpuratus""+""Cephalodiscus gracilis""+""Saccoglossus mereschkowskii""+ ""Ptychodera bahamensis""+""Schizocardium braziliense""+""Branchiostoma floridae""+ ""Botryllus schlosseri""+""Ciona intestinalis""+""Homo sapiens""+""Gallus gallus""+ ""Petromyzon marinus"""	None	None	1
All	"Cliona varians+""Salpingoeca rosetta""+""Xenoturbella bocki""+""Branchiostoma floridae""+ ""Botryllus schlosseri""+""Macrodasys sp""+""Membranipora membranacea""+""Petromyzon marinus""+ ""Eumecynostomum macrobursalium""+""Pomatoceros lamarckii""+""Astrotomma agassizi""+ ""Labidiaster annulatus""+""Macrostomum lignano""+""Peripatopsis capensis""+""Hofstenia miamia""+ ""Cephalodiscus gracilis""+""Phoronis psammophila""+""Adineta vaga""+""Priapulus caudatus""+ ""Lottia gigantea""+""Euplokamis dunlapae""+""Schizocardium braziliense""+""Sycon ciliatum""+ ""Halicryptus spinulosus""+""Ptychodera bahamensis""+""Cephalothrix hongkongiensis""+ ""Leucosolenia complicata""+""Strongylocentrotus purpuratus""+""Diopisthoporus longitubus""+ ""Capitella teleta""+""Leptochiton rugatus""+""Novocrania anomala""+""Crassostrea gigas""+ ""Lepidodermella squamata""+""Helobdella robusta""+""Oscarella carmela""+""Ixodes scapularis""+ ""Mnemiopsis leidyi""+""Stomolophus meleagris""+""Ciona intestinalis""+""Nemertoderma westbladi""+ ""Schmidtea mediterranea""+""Ascoparia sp""+""Strigamia maritima""+""Brachionus calyciflorus""+ ""Convolutriloba macropyga""+""Monosiga brevicollis""+""Terebratalia transversa""+ ""Loxosoma pectinaricola""+""Meara stichopi""+""Lineus longissimus""+""Hemithiris psittacea""+ ""Childia submaculatum""+""Adineta ricciae""+""Daphnia pulex""+""Taenia pisiformes""+ ""Prostheceraeus vittatus""+""Megadasys sp""+""Nematostella vectensis""+""Schistosoma mansoni""+ ""Eunicella cavolinii""+""Drosophila melanogaster""+""Isodiametra pulchra""+""Leptosynapta clarki""+ ""Homo sapiens""+""Barentsia gracilis""+""Craspedacusta sowerby""+""Gallus gallus""+ ""Amphimedon queenslandica""+""Saccoglossus mereschkowskii""+""Trichoplax adhaerens""+ ""Diopisthoporus gymnopharyngeus""+""Dumetocrinus sp""+""Pleurobrachia bachei""+ ""Aphrocallistes vastus""+""Acropora digitifera""+""Agalma elegans""+""Sterreria sp"""	None	None	0
Protostomia	Spiralia+Ecdysozoa	None	Spiralia+Ecdysozoa	1
Nephrozoa	Protostomia+Deuterostomia	None	Protostomia+Deuterostomia	1
Bilateria	Nephrozoa+Xenacoelomorpha	None	Nephrozoa+Xenacoelomorpha	1
Xenambulacraria(C1)	Ambulacraria+Xenacoelomorpha	None	None	1
C2	Chordata+Xenambulacraria(C1)	None	None	1
Bilateria(C3)	C2+Protostomia	None	None	0
D2	Xenacoelomorpha+Deuterostomia	None	None	1
Bilateria(D3)	D2+Protostomia	None	None	0
Xenambulacraria(E1)	Xenoturbella+Ambulacraria	None	None	1
E2	Chordata+Xenambulacraria(E1)	None	None	1
E3	Protostomia+E2	None	None	1

Figure 6.7: Complete split definitions for the Xenoturbella dataset of [43].

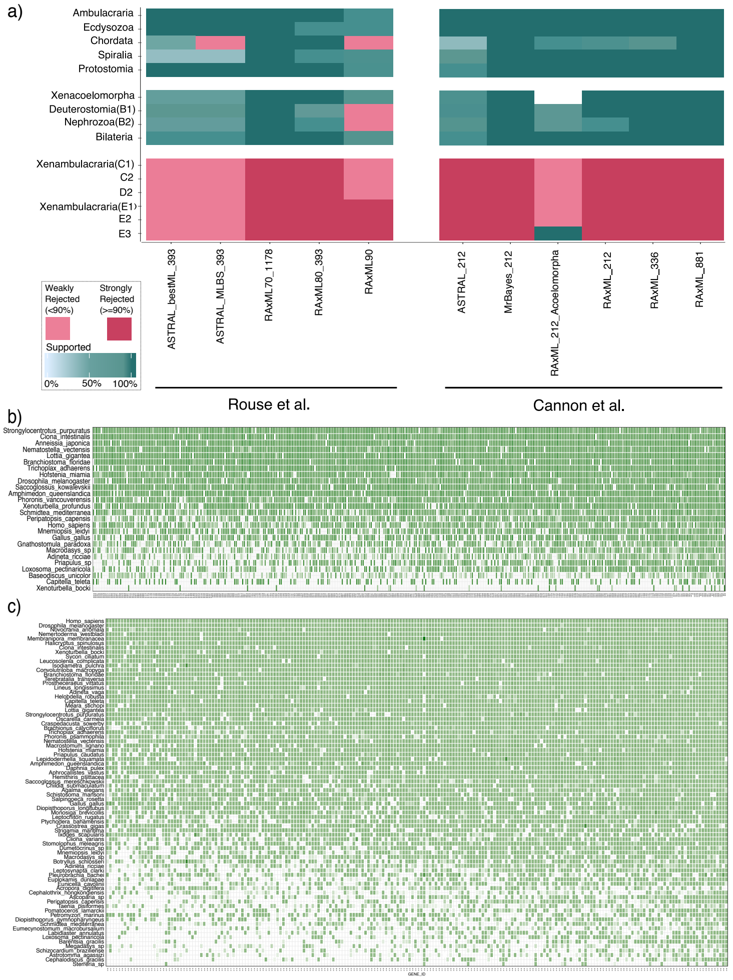


Figure 6.8: Species tree discordance and occupancy. a) DiscoVista Specie tree analysis. Rows correspond to focal splits, and columns correspond to alternative species trees reported in two papers [203, 43]. The spectrum of blue-green indicates support values for splits compatible with a tree. Note that support values are of different types (Bayesian posterior, concatenation bootstrap support, and multi-locus bootstrap support) and thus may not be directly comparable. Weakly rejected splits correspond to splits that are not present in the tree, but are compatible if low support branches (below 90%) are contracted. In [203], RAxML70_1178, RAxML80_393, and RAxML90 are results of concatenation analyses on genes with average occupancy of 70%, 80%, and 90% respectively. For [43], RAxML_212, RAxML_212_Acoelomorpha, RAxML_336, and RAxML_881 are concatenation analyses on 212 orthologus genes of 78 species, 212 orthologus genes after removing Acoelomorpha, 336 orthologus genes on 56 selected species, and 881 orthologus genes on 77 species, respectively. The ASTRAL_212 is the species tree estimated using ASTRAL and 212 orthologus genes of 78 species. b,c) The occupancy map of 393 (b) and 212 (c) gene alignments of [203] (b) and [43] (c) with average occupancy of 80% in both datasets. The spectrum of green color shows the gene length, and the white color indicates missing data.

with the tree when branches with support values below the input threshold (e.g., 75% BS) get contracted. Compatibility is a concept that I find useful for measuring discordance in an interpretable way. A split is compatible with an unrooted tree if it is compatible with all splits of that tree and can therefore be added to that tree; compatibility can be easily and efficiently checked [260]. Let two splits be $A|B$ and $C|D$; then, the two splits are compatible if and only if one of the four pairwise intersections $A \cap C$, $A \cap D$, $B \cap C$, or $B \cap D$ is empty.

On the empirical datasets, the species tree compatibility figures (Fig. 6.8a) show several patterns. The two datasets produce largely congruent results, with only minor differences. In the [203] dataset, aggressive filtering of genes based on their occupancy negatively impacts concatenation trees in terms of BS support and the recovery of some clades (e.g., Chordata). Consistent with literature [160], ASTRAL run on maximum likelihood gene trees (bestML) seems more accurate than ASTRAL run on the bootstrap replicates (MLBS) gene trees (impacting the recovery of Chordata).

ii) Occupancy analysis Missing data is common in phylogenomics, and common filtering practices can further increase the amount of missing data with potential impact on downstream analyses (e.g., [97, 148]). DiscoVista can visualize the taxon occupancy for individual species or for groups of taxa; taxon occupancy is defined as the fraction of gene alignments that have at least one of the species from a group. The inputs to this analysis are a set of sequence alignment files (FASTA format), and a species annotation file; the output is a line plot of the occupancy per species or per group. DiscoVista can also visualize taxon occupancy as a heatmap that shows the presence/absence of species in gene alignments, and for those present, uses color shades to indicate their non-gapped sequence length compared to the other sequences of the same gene.

On the empirical [203] dataset, the occupancy plots immediately reveal how *Xenoturbella bocki* has low occupancy (Fig. 6.8b), appearing only in a handful of genes. However, the negative impacts of this low occupancy may be ameliorated by the high occupancy of the other two Xenacoelomorpha species, *Xenoturbella profoundus* and *Hofstenia miamia*. Patterns of occupancy do not vary widely on this dataset. On the Cannon dataset, in contrast, some genes have lower levels of occupancy than others (from left to right in Fig. 6.8c). Just like the Rouse dataset, several species have very low occupancy (e.g., *Astrotomma agassizi, Cephalodiscus gracilis, Sterreria sp.*). However, *Xenoturbella bocki*, which is the only Xenacoelomorpha species in this dataset, has high occupancy and appears in most of the genes.

iii) Gene tree compatibility This visualization simply depicts the portion of gene trees supporting or rejecting each focal split. The inputs are one or several collections of gene trees, the split definition file, and a support threshold. For splits that are compatible with gene trees, branches with bootstrap support values above (below) a threshold are considered as highly (weakly) supported. Splits that are not compatible with the original tree but are compatible with the tree if branches with low support values get contracted are considered as weakly rejected splits, and those that are not compatible even after contracting low support branches are considered strongly rejected. By default, gene trees that miss one side of the split completely are removed from the analysis. The figure can also be created such that genes that completely miss one side of the split (or completely miss a user-defined subset of a side of the split) are marked as "missing" (Fig. 6.10).

On the empirical datasets, these figures reveal high levels of gene tree discordance (Fig. 6.9a). Although ASTRAL species trees on maximum likelihood gene trees are mostly congruent in both datasets and with concatenation results, gene trees show high amounts of discordance, and none of the major splits are recovered in most gene trees (Fig. 6.9a). However, for all of the focal splits, the majority of the gene trees are compatible with the species tree after contracting low support branches (below 75%). Interestingly, maximum likelihood gene trees in [203] show somewhat less discordance with major species tree splits compared to those in [43].

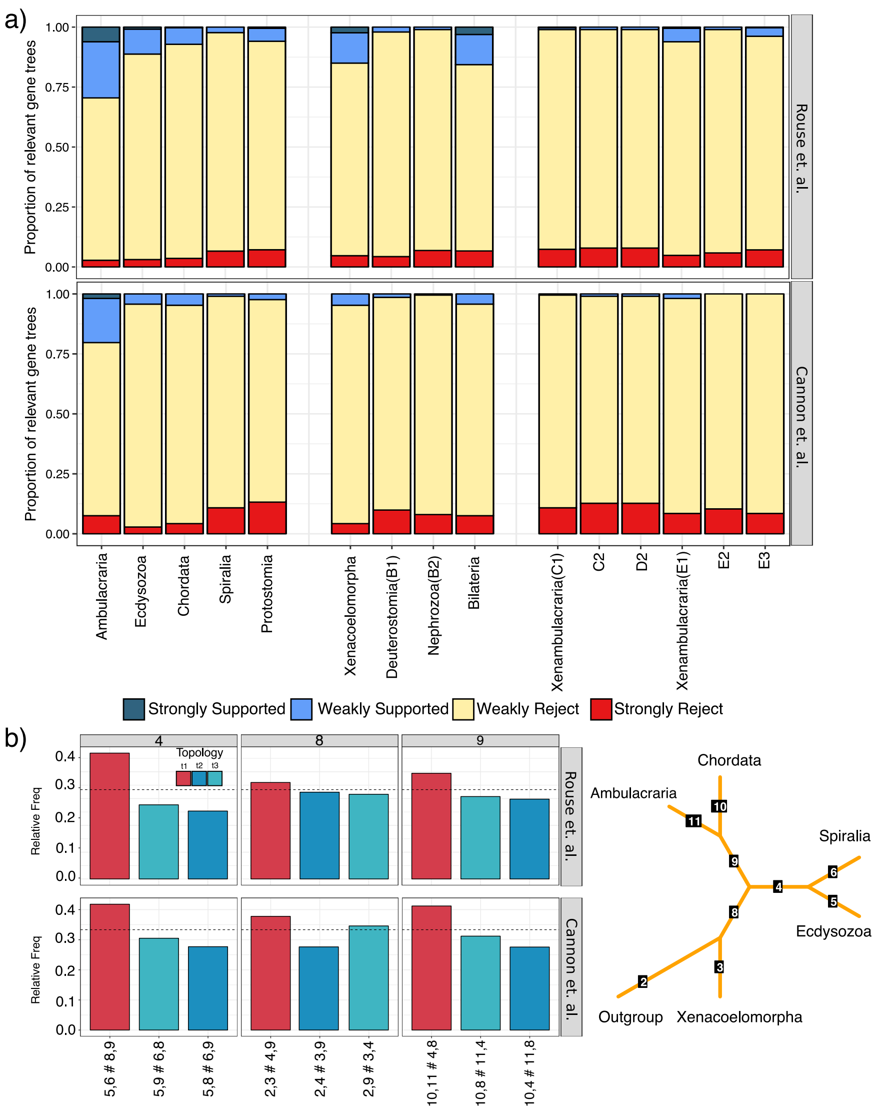


Figure 6.9: Gene tree discordance figures. a) Gene tree compatibility. This figure shows the portion of RAxML genes for which focal splits (x-axis) are highly (weakly) supported or rejected. Weakly rejected splits are those that are not in the tree but are compatible if low support branches (below 75%) are contracted. b) Frequency of three topologies around focal internal branches of ASTRAL species trees in both datasets. Main topology is shown in red, and the other two alternative topologies are shown in blue. The dotted line indicates the 1/3 threshold. The title of each subfigure indicates the label of the corresponding branch on the tree on the right (also generated by DiscoVista). Each internal branch has four neighboring branches which could be used to represent quartet topologies. On the x-axis the exact definition of each quartet topology is shown using the neighboring branch labels separated by “#”.

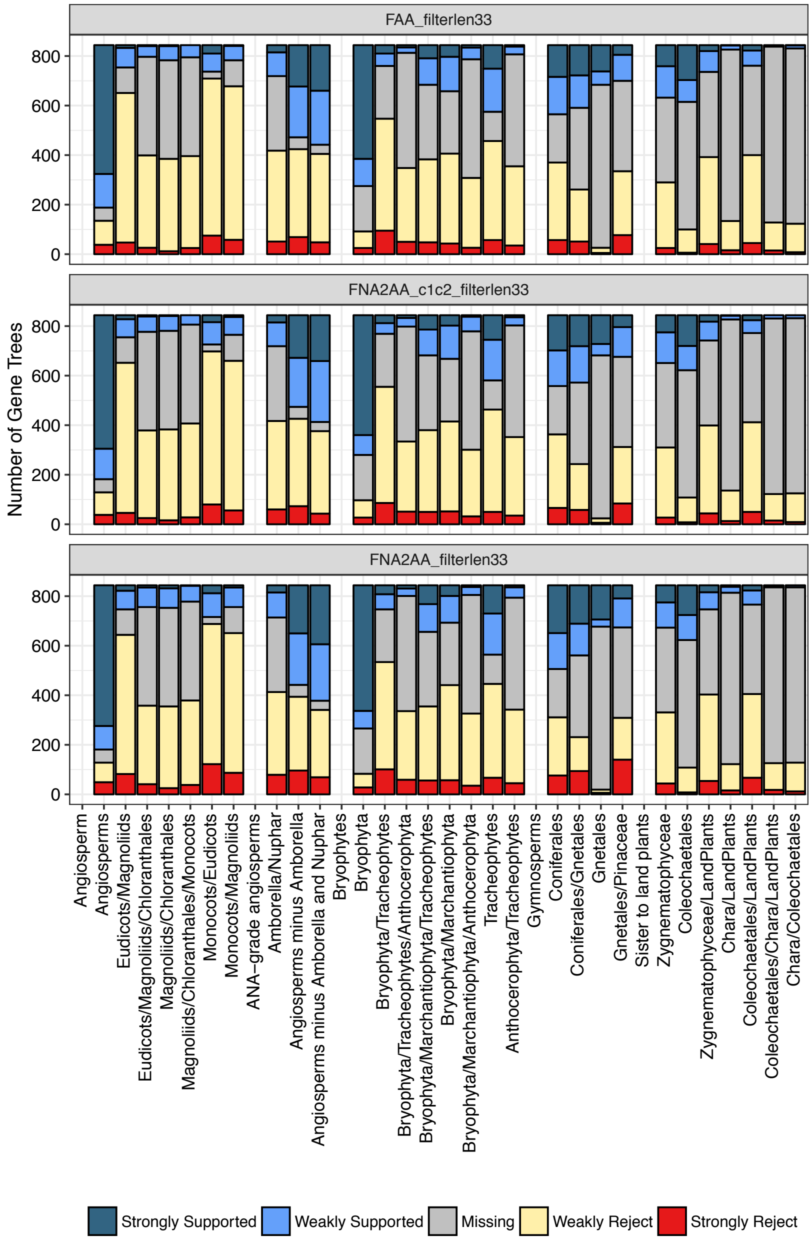


Figure 6.10: Gene tree analysis on 1kp dataset: The portion of RAxML genes for which important clades (x-axis) are highly (weakly) supported or rejected for three model conditions of the 1kp dataset. FAA-filterlen33: gene trees on amino acids sequences, and fragmentary sequences removed (66% gaps or more) FNA2AA-f25: amino acid sequences back translated to DNA, and sequences on long branches (25X median branch length)removed; FNA2AA-filterlen33: amino acid sequences back translated to DNA, and fragmentary sequences removed (66% gaps or more). Weakly rejected clades are those that are not in the tree but are compatible if low support branches (below 75%) are contracted

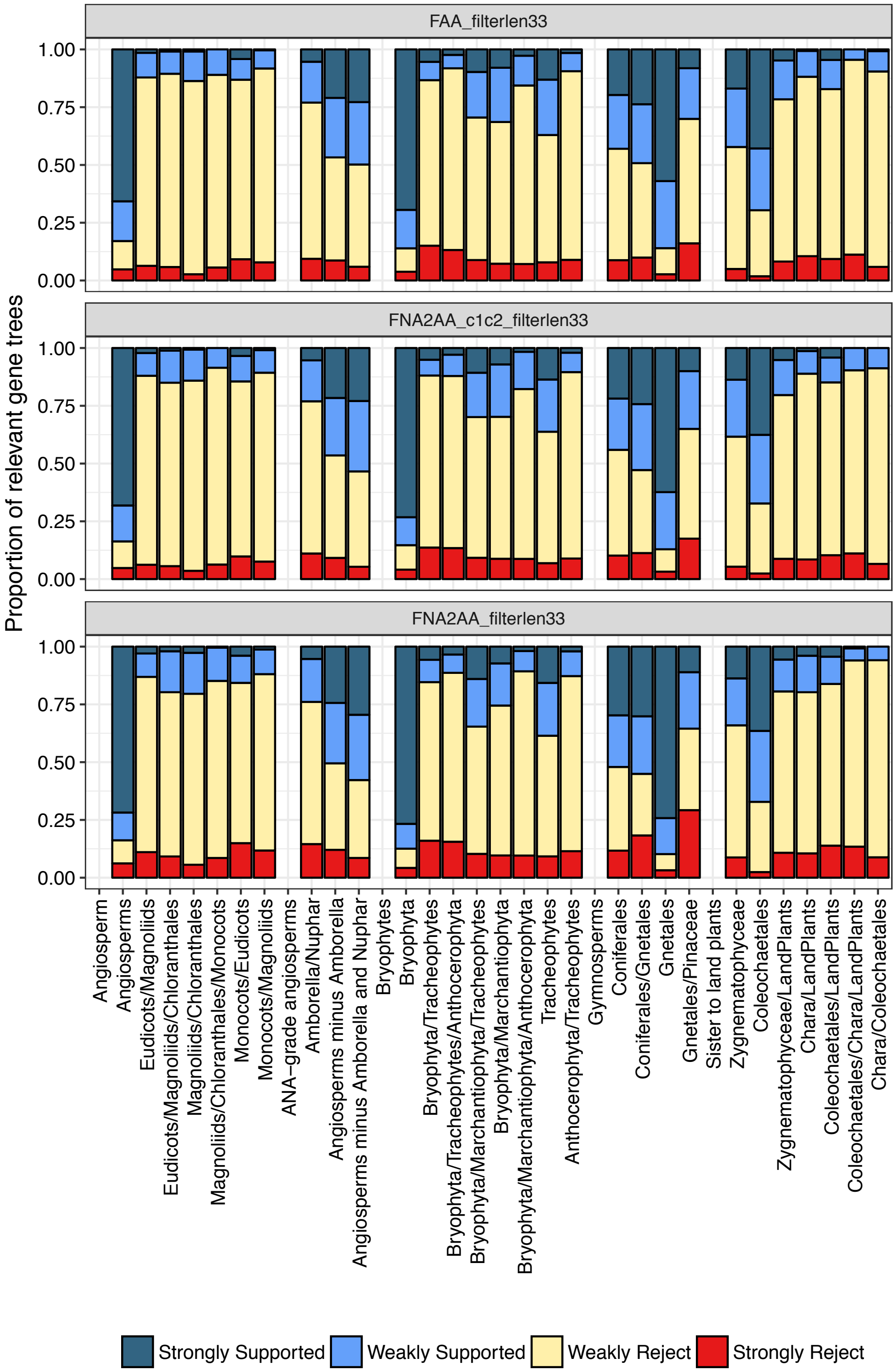


Figure 6.11: Gene tree analysis on 1kp dataset: The number of RAxML genes for which important clades (x-axis) are highly (weakly) supported or rejected or are missing of three model conditions (same as 6.10) of the 1kp dataset. Weakly rejected clades are those that are not in the tree but are compatible if low support branches (below 75%) are contracted.

iv) Branch quartet frequencies Every internal branch of a species tree divides the tree into four parts, and thus, at least one quartet tree (often many) can be mapped uniquely onto that branch. For a given branch, assuming (a) the branch is correct (b) the only source of discordance between gene trees and species tree is ILS, and (c) there is no gene tree estimation error, the multi-species coalescent (MSC) model has specific expectations about these quartets [7]. The probability of a gene tree quartet tree matching the species tree topology ($p > \frac{1}{3}$) is higher than probability of matching the two alternatives, and the two alternatives have equal probabilities ($q = \frac{1-p}{2} < \frac{1}{3}$). Visualizing the empirical frequency of quartets around each branch can serve two purposes: it gives an interpretable measure of discordance specific to that branch [210], and one can check whether the assumptions of ILS are met. If the discordance is purely due to ILS, then one would expect the second and the third hypotheses to have similar frequencies (close to q). Lack of this pattern can have many causes, including hybridization [225]. Finally, note that if the length of a species tree branch in coalescent units [7] is d, then for quartets around it, $p = 1 - \frac{2}{3}e^{-d}$, and thus, the quartet frequencies also convey information about the branch length. In the limit, for $d = 0$ (e.g., a polytomy) one would expect all three frequencies to be close to $\frac{1}{3}$ and $p = q = \frac{1}{3}$ can be tested as a null hypothesis [211].

In this visualization, for every focal split, a bar graph shows the quartet frequencies around that branch. Moreover, a cartoon tree is generated where leaves are the large groups of taxa collapsed into single leaves and branches are labeled consistently with the bar graph. Inputs to this analysis are a set of gene trees, a species tree, an annotation file that maps each species to a named group and thus defines leaves of the cartoon tree and by extension, the focal splits.

The portion of quartets around three focal branches of the ASTRAL species trees in my empirical studies show why some branches have been controversial (Fig. 6.9b). For Nephrozoa (branch labeled as “8”) and Deuterostomia (labeled as “9”), the relative frequency of main topologies are extremely close to 1/3 in both studies. This high level of gene tree discordance around these two branches is likely caused by a combination of high true discordance and also gene tree estimation error; irrespective of which is more prevalent, the high discordance reveals a cause of difficulties faced in resolving these relationships. Interestingly, the portion of quartets for alternative topologies in [203] follow the expectations of the MSC theory quite well (i.e., second and third topologies have very close frequencies); the same cannot necessarily be said of the [43] gene trees.

v) GC content Commonly used models of sequence evolution are stationary and assume that all species have identical base composition. This assumption is often violated in gene coding sequences, especially in the third codon position [106]. While non-stationary models [32] and tests of divergence from stationarity [2] exist, simply visualizing GC content of different codon position for different extant taxa can help in judging non-stationarity and in deciding whether a GTR analysis is appropriate for some or all of the data [105, 264]. DiscoVista generates box plots (distributions) or dot plots (averages) of GC content of each codon position, as well as all three codon positions, for each species. The input to this analysis is a set of gene coding sequences in FASTA format.

The two Xenacoelomorpha datasets did not release DNA sequences and therefore, I could not compute their GC content. Nevertheless, examples of the GC content figures could be seen for the plant dataset (Fig. 6.5). These figures immediately show that assumptions of equal base frequencies are violated. They also show that the variations in the GC content are mostly concentrated in the third codon position. The GC content levels for the second codon position, and to a lesser degree the first codon position, do not vary much across species. These results favor the removal of the third codon position when building gene trees or in concatenation analyses.

To summarize, DiscoVista provides a useful tool for visualizing several patterns of discordance, missing data, and GC variations in phylogenomic datasets.

6.3 Acknowledgements

Chapter 6, in full, contains material from SAYYARI, E., WHITFIELD, J. B., AND MIRARAB, S. DiscoVista: Interpretable visualizations of gene tree discordance. Molecular Phylogenetics and Evolution 122 (2018), 110–115. I was the primary investigator and author of this paper.

Chapter 7

TADA

Learning associations of traits with the microbial composition of a set of samples is a fundamental goal in microbiome studies. Recently, machine learning methods have been explored for this goal, with some promise. However, in comparison to other fields, microbiome data is high-dimensional and not abundant; leading to a high-dimensional low-sample-size under-determined system. Moreover, microbiome data is often unbalanced and biased. Given such training data, machine learning methods often fail to perform a classification task with sufficient accuracy. Lack of signal is especially problematic when classes are represented in an unbalanced way in the training data; with some classes under-represented. The presence of inter-correlations among subsets of observations further compounds these issues. As a result, machine learning methods have had only limited success in predicting many traits from microbiome. Data augmentation consists of building synthetic samples and adding them to the training data and is a technique that has proved helpful for many machine learning tasks. In this dissertation, I propose a new data augmentation technique for classifying phenotypes based on the microbiome. My algorithm, called TADA, uses available data and a statistical generative model to create new samples augmenting existing ones, addressing issues of low-sample-size. In generating new samples, TADA takes into account phylogenetic relationships between microbial species. On two real datasets, I show that adding these synthetic samples to the training set improves the accuracy of downstream classification, especially when the training data have an unbalanced representation of classes

7.1 Introduction

Understanding the impact of the composition of the microbiome on clinically-relevant traits is a major promise of microbiome profiling [172] using both 16S [79] and metagenomic sampling [256]. The goal is to understand how the composition of species, or genes, in a microbial community such as human gut impacts phenotypes of interest such as obesity (e.g., [254]). The relationship between microbial composition and traits, however, is complex and hugely variable, from person to person [52] and from one time to another [44, 73]. As a result, microbial communities have been hard to model [258] using traditional sample differentiation methods [181, 123].

Machine learning (ML) methods have proved capable of capturing complex relationships in many fields, such as vision and speech recognition. As a result, researchers have pointed out the potential of ML models to capture complexities of the microbiome [114]. Many researchers (e.g., [233, 208]) have formulated understanding microbiome as a classification task: given is a set of samples, each consisting of a set of sequences from various microorganisms, and each sample is labeled by a trait of interest (e.g., lean or obese); a model is learned to predict these labels and classify unlabeled (new) samples. Some studies have shown promise in achieving an accurate classification of clinically relevant traits using microbiome (e.g., [1, 71, 22]).

The number of samples available for training an ML algorithm has tremendous effects on the accuracy of the model. Tuning a large number of parameters of a classifier or regression method using a small dataset can lead to overfitting and poor generalization to new samples.

Impacts of overfitting are particularly severe when I have an unbalanced distribution of class labels or hidden confounding factors in training datasets (e.g., [121, 46, 45]).

The number of microbiome samples, compared to applications like vision and speech recognition, is relatively small. For example, ImageNet Large-Scale Visual Recognition Challenge (ILSVRC-2010) involved the classification of 1.2 million high-resolution images into 1000 different classes [204], whereas, one of the largest microbiome datasets, the American Gut Project (AGP) [153], includes 14,794 samples, has only self-reported labels, and is heterogeneous (e.g., only 1,942 samples are omnivores of age between 20 and 80 with no self-reported disease or antibiotic usage). For classifying specific traits, AGP has even fewer samples (e.g., only 262 samples report having inflammatory bowel disease). Moreover, the representation of traits of interest is often not balanced, and the distribution of the labels often is not even close to the larger population (e.g., targeted datasets are often over-represented in the diseased state and short on healthy samples). Biases are further compounded by the natural variability of microbiome and auto-correlation between labels due to hidden or nuisance variables, which abound. These difficulties have lead to diminished hope for the generalization of methods [247].

Perhaps the ultimate goal should be gathering more (and less biased) labeled samples for training, a task that will progress only slowly, especially given difficulties of combining datasets gathered with various lab protocols [261, 127]. An alternative that has has been explored extensively in recent years by the ML community is data augmentation. The idea is to create artificial labeled samples algorithmically and add them to the training data. For example, two widely-used methods, SMOTE [46] and ADASYN [89] seek to reduce biases introduced by unbalanced distributions of labels using a k-NN clustering of samples and combining points in the same cluster. Beyond these generic methods, which do not seek to capture domain knowledge, augmentation has the potential to combine the power of black-box ML models and biologically-motivated generative statistical models.

In this dissertation, I propose a new data augmentation technique for microbiome data, called Tree-based Associative Data Augmentation (TADA). The main ideas behind TADA are two-fold. *i*) Each observed sample captures the underlying microbiome only imperfectly, and hence, a variation of the sample could have easily been observed, *ii*) such variations are constrained by the phylogenetic relationships between species [150], which underlie the sequence similarity and microbial diversity [177, 257]. Thus, TADA generates new samples while considering the evolutionary relationships between organisms. Furthermore, I do not stop at just increasing the number of samples. As I will show, it is crucial to deal with unbalances and biases in the training data. In deciding what samples to add, TADA can also remove unbalances in the data with respect to both observed and hidden variables (which I seek to approximate using clustering). I test TADA on two datasets with various biases added to the training dataset. I show that two leading ML models (random forests and neural networks) fail to perform well on unbalanced and biased samples. I also show that data augmentation improves the accuracy, marginally but meaningfully for balanced datasets and dramatically in the presence of unbalanced training sets.

7.2 The TADA method

7.2.1 Background and Notations

The training data used in microbiome classification is an operational-taxonomic-unit (OTU) table **X**. The rows of the table correspond to a set of m samples, often one per individual, $\mathcal{S} = \{s_1, s_2, \ldots, s_m\}$ and the columns correspond to features. Features can be defined in various ways, but for simplicity, I focus on a specific form. My features are a set of n OTUs (e.g., representing species), $\{o_1, o_2, \ldots, o_n\}$. Each cell of the matrix gives the number of times an OTU is observed in a sample. The counts in each row can also be normalized so that they add up to one. In addition to the OTU table, I need a class label y_i for each sample s_i. The class labels correspond to phenotypes (e.g., healthy versus diseased or lean versus obese) that I seek to classify using the microbiome.

The OTUs have a corresponding sequence, for example from the marker genes like 16S rRNA. These sequences may be obtained using a number of approaches, including the traditional OTU picking methods [58, 213] or sub-operational-taxonomic-unit (sOTU) methods [8, 41, 59]. Depending on the method, the exact meaning of OTUs changes; however, they always correspond (at least in approximation) to microorganisms that constitute the sample.

7.2.2 Generative model used in TADA

Data augmentation seeks to add to the training set new samples that could have been seen but are not seen. TADA achieves this using a generative model to create synthetic samples distributed around existing samples. TADA models two types of variations.

- *True variation (TV)*. From one individual to another, even among those with the same phenotype, the true proportions of different OTUs in the microbiome change. These variations may be due to confounding factors (i.e., hidden variables) or natural biological variation among people. Moreover, the microbial composition for each person may also change through time. Thus, samples have true biological variation.

- *Sampling variation (SV)*. Environmental sequencing takes a random (but not necessarily uniformly random) subsample of the true diversity, creating additional variation around the true proportions. Moreover, sequencing adds errors and ambiguity that further increase variation.

Of the two forms of variation, true variation is much harder to model statistically. Confounding

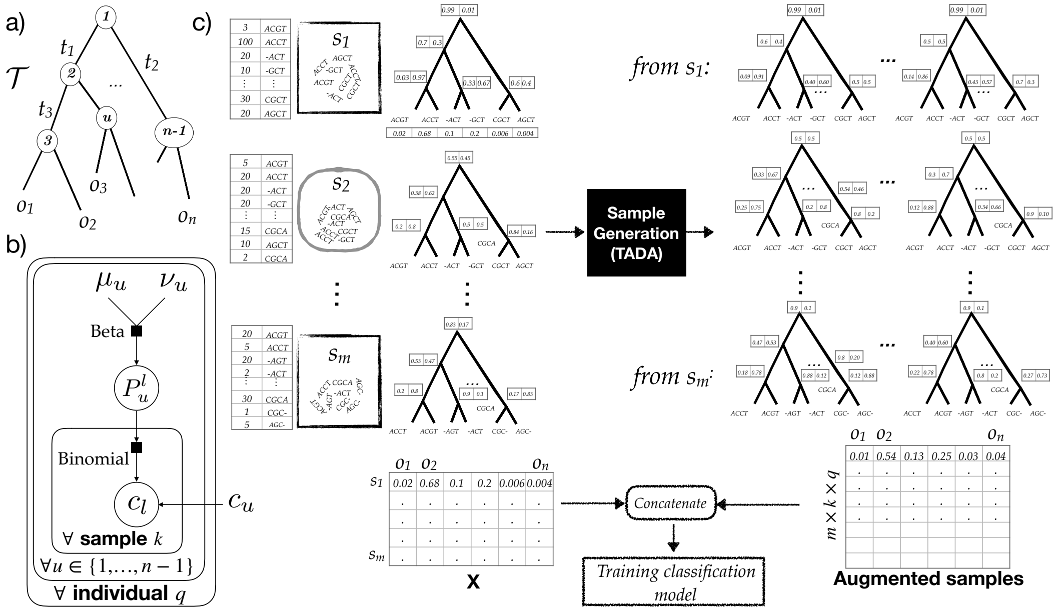


Figure 7.1: (a) A phylogeny $\mathcal{T}$ with branch lengths (t_u), OTUs at leaves (o_i), and internal node indices. (b) The hierarchical graphical model used to generate new samples. (c) The augmentation procedure. First, each sample is mapped to the phylogeny, then I estimate parameters of the model for each sample s_i (or a collection of samples; see Figure 7.2), and then generate new samples using the generative model. The augmented samples are concatenated with the original samples for training the classifier (e.g., RF or NN).

factors are mostly unknown as are the source of natural or temporal variations. However, a major source of inter-correlation, the phylogenetic structure, *can* be inferred and modeled.

Phylogenetic structure. Microorganisms that make up a sample are all descendants from a common ancestor, as captured by their phylogenetic tree. The shared evolutionary history creates a dependence between OTUs, and a phylogenetic tree can represent the relationships (in its topology) as well as the distance between the species. Close phylogenetic relationships between OTUs corresponds to closeness in the sequence space and perhaps also in functional roles. Both forms of variation are likely influenced by the phylogeny. True variation can be phylogenetic because phylogenetically similar organisms may interchange easily, though I note that this is far from a universal rule; strain-variation may have a large impact on the function. Sampling variation is impacted because algorithms for creating OTU tables are prone to merge or confuse OTUs that are close phylogenetically.

TADA uses an inferred binary phylogenetic tree (details are given in Section 7.3.4), called $\mathcal{T}$, with leaves labeled by OTUs $o_1 \ldots o_n$ (Fig. 7.1a). I index internal nodes of $\mathcal{T}$ from 1 (for the root) to $n-1$ and refer to the length of the edge above node u by t_u. Using a simple $O(n)$ algorithm (Algorithm 6), I compute d_u: the average length of the path from each leaf under the left child of u to each leaf under the right child of u.

Algorithm 6 Algorithm to compute the average tip-to-tip distances of tree. Length of each node is defined as the length of the edge above it.

```
procedure AVERAGE_TIP_TO_TIP($\mathcal{T}$)
    for $u \in$ postorder traversal of $\mathcal{T}$ do
        if $u$ is a leaf then
            $u$.num = 1
            $u$.avg = 0
            $u$.sum = 0
        else
            $v, w$ = left and right children of $u$
            $u$.num = $v$.num + $w$.num
            $u$.avg = ($w$.sum $\times$ $v$.num + $v$.sum $\times$ $w$.num)/($v$.num $\times w$.num) $+t_v + t_w$
            $u$.sum = $v$.sum + $w$.sum + $t_v$ $\times$ $v$.num + $t_w$ $\times$ $w$.num
    return $\mathcal{T}$
```

Generative model: the base model

I design a hierarchical generative model to capture both sources of variation and the phylogenetic auto-correlation. The model has three sets of parameters: $i)$ the phylogeny, $\mathcal{T}$, and its branch lengths (and, thus, d_u's), with the two nodes below each node u arbitrarily labeled as *left* (l) and *right* (r), $ii)$ a set $\mathcal{M} = \{\mu_1 \ldots \mu_{n-1}\}$, $0 < \mu_u < 1$, each corresponding to an internal node of the phylogenetic tree, $iii)$ the total sequence count N. In addition to these parameters, I define for each node, a value $\nu_u = f(d_u)$ where f can be any monotonically increasing function.

My generative hierarchical model (Fig. 7.1b) is defined recursively, starting at the root and traversing the tree top-down. Algorithm 7 shows this model generates q individuals and k

new samples for each individual ($k \times q$ in total), each with N sequences. The true variation is modeled using a Beta distribution and the sample variation using a Binomial distribution. I use the μ, ν parameterization of the Beta distribution (as opposed to the standard α, β parameterization). For each node u, I have the parameter μ_u, which gives the population-wide portion of sequences under the node u that fall under the left subtree of u. A draw from the Beta distribution gives us p_u^l: the true portion of sequences that go to the left subtree in the underlying microbiome. Then, a draw from the Binomial distribution gives the actual observed count and models the variation due to sampling (sequencing) around the true proportion p_u^l.

In this model, the true variance is inversely proportional to the square root of phylogenetic distance. In the parameterization of the Beta distribution used here, the mean is μ and the variance is $\frac{\mu(1-\mu)}{\nu+1}$. By setting the ν parameter of Beta to a monotonically increasing function of d_u, I make sure that the variance increases closer to the tips of the tree (where d_u is small), and decreases towards the root (where d_u is high). The choice of the exact function f (see Section 7.3.4) is arbitrary. However, the fact that variance should be higher closer to tips has a biological justification. Closer to the leaves of the tree, microbial organisms become more similar and therefore more likely to be able to replace each other in an environment or be confused with each other. Conversely, the microbial composition becomes more stable close to the root of the tree.

Algorithm 7 TADA sample generation procedure.

```
for individual $1 \leq i \leq q$ do
    for node $u$ in preorder traversal of $\mathcal{T}$ do
        Draw $p_u^l \sim Beta(\mu_u, \nu_u)$
    for $1 \leq j \leq k$ do
        $c_1 \leftarrow N$                                  % Index 1 refers to the root node
        for internal node $u$ with children $l$ and $r$ in preorder traversal do
            Draw $c_l \sim Binomial(p_u^l, c_u)$
            $c_r \leftarrow c_u - c_l$
        Output $c_{o_1}, \ldots, c_{o_n}$ as a new sample and normalize if needed.
```

Generative model: mixtures

The model described above is limited in a fundamental way: it assumes all samples are generated from the same underlying distribution. Therefore, it completely ignores the fact that individuals belong to several classes (the identification of which is the goal) and that within each class, confounding factors may create further structure among samples. For example, I may have healthy and diseased samples for my main classes, and for each of those, samples may be further differentiated based on age, gender, weight, or other factors (which, may not be known). Thus, the phenotype structure of samples is not modeled.

To capture the phenotype structure, I use a mixture model. The population is assumed to be divided into clusters, each with its own $\mathcal{M}$ parameters, but all sharing the same phylogeny. Clusters can correspond to class labels, confounding factors, or a mixture of the two. In the generative process, each sample is first assigned to a cluster, according to cluster probabilities, and then the procedure described above is followed.

7.2.3 Data augmentation procedure

Assuming the training data come from my generative model, I can design parameter estimators and use the estimated parameters to generate new data. In fact, a model-based approach (coupled with the mixture model), in principle can also infer the class labels. However, on typical microbiome training datasets, the total number of mixture components is likely large, and the model has a large number of parameters. Thus, parameter estimation using these complex models will be underpowered. Moreover, despite a large number of parameters, the model does not come close to capturing all the biological complexity of microbiome. Thus, instead of using the generative model for inference, I use it only as a tool for data augmentation for training ML models.

Based on the hierarchical model, I design two versions of TADA, which vary in their ambition, ranging from capturing only sampling variation to capturing both sources of variation and confounding factors. The more ambitious versions include more parameters, and this reliance on more parameters makes them vulnerable when applied to limited datasets.

TADA-SV: This version only captures sampling variation and has a single user setting: a number k. For each training sample s_i, I first estimate p_u^l in my training set *independently* from other samples (assuming samples are unlinked). Then, for each sample, k new samples are generated and added to the training set using the fixed p_u^l following Algorithm 7 (setting $q = 1$ and starting in line 4). Thus, this method is only drawing from the Binomial component of my Hierarchical model and ignores the rest. To estimate p_u^l from a single sample s_i, I use the total count of sequences that fall to the left of the node u in s_i, normalized by the total count of sequences below u. As proved in Lemma 6, this estimator gives the joint ML estimate for all p_u^l values (treated as parameters of the binomial) over the entire tree.

Lemma 6. Consider the phylogeny $\mathcal{T}$ on the OTU set $\mathcal{S}$. For each internal node u of $\mathcal{T}$, let C^u be random variable giving the number of observed sequences contained under the node u and let c^u be the observed value of C^u for one sample. Assume $C^{l(u)} \sim Bin(p_u, C^u)$ where where $l(u)$ is the left child of u and let $C^{r(u)} = C^u - C^{l(u)}$ where $r(u)$ is the right child of u. The joint maximum likelihood estimate of all p_u values is given by

$$(\hat{p}_1, \dots, \hat{p}_u, \dots, \hat{p}_{n-1}) = (\frac{c^{l(1)}}{c^1}, \dots, \frac{c^{l(u)}}{c^u}, \dots \frac{c^{l(n-1)}}{c^{n-1}}) \,. \tag{7.1}$$

Proof. Consider a sample with the count vector $X = (x_1, x_2, \dots, x_n)$ at the leaves of the tree. Recall the root is indexed 1 and thus $\sum_{i=1}^n x_i = c^1$. Let $path_v^u$ indicate the path from the leaf node v to the node u. The likelihood of observing the count vector $x = (x_1, x_2, \dots, x_n)$ given the phylogeny $\mathcal{T}$, and the conditional probability vector $p = (p_1, \dots p_u, \dots, p_{n-1})$ equals

$$\begin{aligned}\mathcal{L}(X = (x_1, x_2 \dots, x_n); \mathcal{P} =&\{(p_1, \dots, p_u \dots, p_{n-1}, p_n, \dots, p_{u+n-1}, \dots, p_{2n-2}, \mathcal{T}) = \\ &\Gamma(X, c^1) \prod_{i=1}^{n} (\prod_{k \in path_i^{root}} p_k)^{x_i}\end{aligned} \tag{7.2}$$

where $p_{n+e-1} = 1 - p_e$, and $\Gamma(X, c^1)$ is a normalization function that doesn't depend on the conditional probability vector p. Consider the left child of the root, $l(1)$, and recall the probability of sequences falling below it is p_1. All the root-to-leaf paths descending from $l(1)$ have the branch connecting the root to $l(1)$; thus, for these, the probability p_1 is multiplied each time. A similar argument works for the right child of the root with the probability $1 - p_1$. So the likelihood could be written as

$$\begin{aligned}\mathcal{L}(X; \mathcal{P}, \mathcal{T}) =&\Gamma(X, c^1)[\prod_{i=1}^{n} (\prod_{k \in path_i^{lr(1,i)}} p_k)^{x_i}].(p_1)^{(\Sigma_{i \in a(l(1))} x_i)}.(1 - p_1)^{(\Sigma_{i \in a(r(1))} x_i)} = \\ &\mathcal{L}(x; \mathcal{P}, \mathcal{T}^l).\mathcal{L}(x; \mathcal{P}, \mathcal{T}^r).(p_1)^{c^{l(1)}}.(1 - p_1)^{c^1 - c^{l(1)}}\end{aligned}$$

where $lr(1,i) = l(1)$ for leaves under the left child of 1 and $lr(1,i) = r(1)$ for leaves under the right child of 1, $a(u)$ is the set of leaves under the node u, and $\mathcal{P}^l$ and $\mathcal{P}^r$ indicate the left and right subtrees of the root. This means I could compute the $\mathcal{L}(x; \mathcal{P}, \mathcal{T})$ as the product of the likelihood of the left subtree, the likelihood of the right subtree of the root, and the probability of observing a total of $c^{l(1)}$ counts for the $\mathcal{T}^l$.

Note that p_1 (same is true for $1 - p_1$) does not contribute to the likelihood of $\mathcal{T}^l$ or $\mathcal{T}^r$ and, hence, to find p_1 I could consider $\mathcal{L}(x; \mathcal{P}, \mathcal{T}^r).\mathcal{L}(x; \mathcal{P}, \mathcal{T}^l)$ as a constant and ignore. Instead of maximizing the $(p_1)^{c^{l(1)}}.(1 - p_1)^{c^1 - c^{l(1)}}$, I could maximize log of this function

$$f(c^{l(1)}, c^1; p_1) = c^{l(1)} \log(p_1) + (c^1 - c^{l(1)}) \log(1 - p_1)$$

and, hence

$$\hat{p}_1 = \frac{c^{l(1)}}{c^1}.$$

This gives us $\hat{p}_1$ for the edges descending from the root. I can use the same argument recursively and compute other probabilities as

$$\hat{p}_u = \frac{c^{l(u)}}{c^u} .$$

$\square$

Corollary 10. Consider Lemma 6, and the likelihood of observing the count vector $x = (x_1, x_2, \dots, x_n)$ given the phylogeny $\mathcal{T}$, and the conditional probability vector p on all leaves (left subtrees) of the phylogeny $\mathcal{T}$. Equation 7.2 is the likelihood of a multinomial distribution with probabilities $p^{leaf} = (p_1^{leaf}, p_2^{leaf}, \dots, p_n^{leaf})$. In this equation

$$p_i^{leaf} = \prod_{k \in path_i^{root}} p_k$$

.

TADA-TVSV-*C*: This version captures both sampling and true variation and optionally also confounding factors. The method has three user settings: k, q, and C. I first cluster samples $s_1 \dots s_m$ into C groups *per classification label* based on the training data $\mathbf{X}$ using any clustering method of choice (see my default choice in Section 7.3.4). These clusters correspond to components of the mixture model I described before; note that instead of using a complex parameter-rich model-based inference of mixture components, I use a clustering method to approximate the components. The hope is that the clustering based on $\mathbf{X}$ captures the hidden phenotype structure, at least partially. The choice of C controls the level of complexity and therefore the number of parameters. For example, acknowledging the difficulty of finding the phenotype structure, I explore the extreme setting of $C = m$ where each sample in my training set belongs to its cluster and therefore is unlinked from others, just like SV. I also explore other settings of C, including $C = 1$.

After clustering, I first estimate $\mathcal{M}$ parameters *per each cluster* using a method of moments. The estimator, as shown in Lemma 7, simplifies to computing the sum of counts on the left child of each node u across all samples of the cluster, normalized by the sum of the counts under the node u. Then, for each cluster, I generate q new individuals and k new samples per individual (thus, $k \times q$ in total). To do so, I follow the generative procedure given in Algorithm 7.

Lemma 7. Consider the phylogeny $\mathcal{T}$ on the OTU set $\mathcal{S}$ and samples $s_1, \dots, s_w$. Let the total number of sequences in each sample be $\mathcal{C} = \{c^1_1, c^1_2 \dots, c^1_w\}$. Assume that the probability of observing a sequence from a species under the left subtree of the node u follows a beta distribution $p^l_u \sim Beta(\mu_u, \nu_u)$ where ν_u is a fixed parameter which depends only on the phylogeny, $\mathcal{T}$, and is therefore given, and $\mu_u \in \mathcal{M}$ is a parameter shared between all samples. Assume that the number of observed sequences contained under the left subtree of u given p^l_u follows a binomial distribution $C^{l(u)} \sim Bin(p^l_u, c^u)$. Then, the method of moments estimate for μ_u is

$$\mu_u = \frac{\sum_{j=1}^{w} c_j^{l(u)}}{\sum_{j=1}^{w} c_j^u} \tag{7.3}$$

where $l(u)$ is the left subtree of u, and c^u_j is the number of observed sequences contained under u in the sample s_j.

Proof. Consider the new random variable $\sum_{j=1}^{w} C_j^{l(u)}$

$$\mathbb{E}[\sum_{j=1}^{w} C_j^{l(u)}] = \sum_{j=1}^{w} \mathbb{E}[C_j^{l(u)}] = \sum_{j=1}^{w} \mathbb{E}_{p^l_u}[\mathbb{E}[C_j^{l(u)} | p^l_u]] =$$
$$\sum_{j=1}^{w} \mathbb{E}_{p^l_u}[c^u_j p^l_u] = \mu_u \sum_{j=1}^{w} c^u_j$$

and hence having

$$\mu_u = \frac{\sum_{j=1}^{w} c_j^{l(u)}}{\sum_{j=1}^{w} c_j^{u}}$$

□

7.2.4 Balancing

So far, I have generated a fixed number of new samples per input training sample. However, by generating a *different* number of samples per input sample, I can use augmentation for balancing (or otherwise adjusting) my input training set in terms of the distribution of labels. As I will show, lack of balance between representation from different phenotype classes (e.g., the training labels) can degrade the accuracy of ML methods. TADA, therefore, can also be run with balancing (Fig. 7.2). In this mode, training data are first divided into several groups; these groups can be based on classification labels, the result of clustering training points, or a combination. I then choose the number of extra samples generated per sample (e.g., k and q) such that all groups have the same number of samples after augmentation. I will test two modes.

- TADA-Balance adds exactly as many new samples as necessary (and not any more) so that all groups have the same total number of samples.

- TADA-Balance++ not only makes all groups balanced in size but also increases the total number of samples for all groups, so that the largest group has q times more samples than before augmentation.

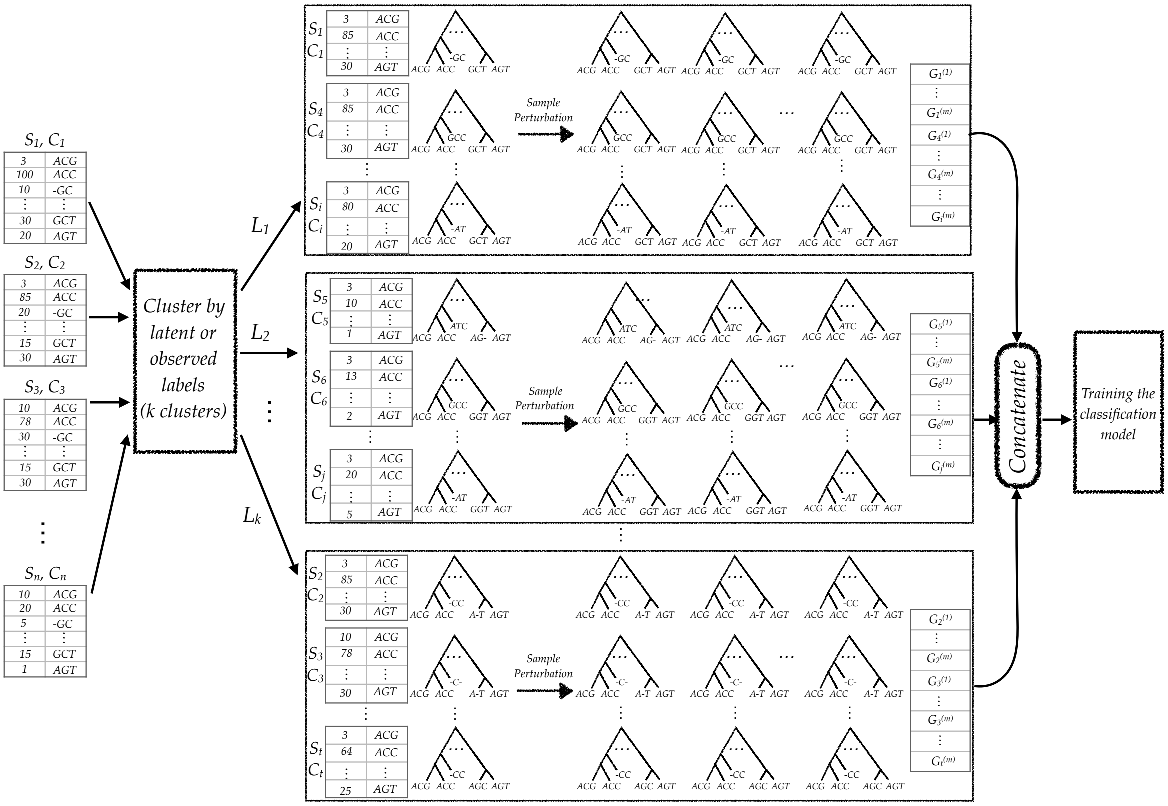


Figure 7.2: In the clustering scenario, I first group samples using k-means clustering applied to the Bray-curtis distances between samples. I then generate new samples for each group separately.

7.3 Experimental setup

7.3.1 Datasets

I use two datasets, both based on 16S profiling of gut microbiome.

Gevers. As my main dataset, I use a dataset by [77] (publicly available on Qiita [80]; study ID 1939), which the authors put together to study the impact of the microbiome on the Inflammatory Bowel Disease (IBD). This study has 1,359 samples, has been gathered in a clinical setting, is carefully curated, and has reliable class labels. I filtered out samples from people on antibiotics, with less than 10,000 16S sequences. Before running my experiment, I also removed 9 outliers and any OTUs with total counts across all samples below 3. This leaves us with 647 diseased samples and 243 healthy samples, gathered using either biopsy or stool. [77] were able to find a clear indication that IBD changes the microbiome composition, and thus, ML methods should be able to achieve reasonable classification accuracy on this dataset.

BMI. In addition, I use the American Gut Project (AGP) [153]. This dataset has only self-reported labels and is gathered by crowdsourcing instead of a clinical setting. Thus, it is less curated than the Gevers dataset, though authors have taken several quality control steps. With an understanding of the shortcomings, I use the AGP data to test my method on a phenotype other than IBD. I classified the self-reported body mass index (BMI) phenotype categorized into 1,360 lean vs. 582 overweight samples (cutoff at BMI: 25). Similar to Gevers dataset, I further filtered this dataset to control for many factors that might affect microbiome composition. These factors include diet (I keep omnivore samples), ethnicity (Caucasian), country (USA), disease (healthy), antibiotics (no antibiotic usage in the past year), and age (between 20 to 70). I also filtered samples with less than 1,000 de-noised reads, one outlier, and I removed OTUs with total counts across all samples below 4.

7.3.2 Experiments

E1. On both Gevers and AGP datasets, I compare TADA, in its two settings, against ADASYN and SMOTE, and the baseline approach where no augmentation is performed. In this experiment, I use all the data for training and testing in a cross-validation setting (see evaluation procedure below) performed using the held-out original samples. **E2.** I next test the impact of balancing by generating unbalanced training data. I created datasets such that 1/10, 1/5, or 1/3 of both the training and testing data are from healthy/overweight individuals. I created two versions of this unbalanced dataset. In the first version (E2-fix), I used a fixed number (243 for IBD and 582 for BMI) training samples for all three ratios to test the impact of the ratio without changing the training size. Here, the testing sets are chosen to match the ratio in the training set but have a larger total sum (the maximum possible in each case). I also created a version (E2-max) only for the IBD dataset with the maximum possible training size for each ratio. To do this, I removed the minimum possible number of samples from healthy (for 1/10 and 1/5 ratios) or diseased (for 1/3) so that I obtain the desired ratios; this leaves us with 574 training samples for 1/10, 646 for 1/5, and 587 for 1/3. Once again, the testing sets are chosen to match the training set in ratio. E2-max enables us to make sure results on E2-fix hold if training datasets are as large as possible. Here, in addition to no-augmentation, I compare TADA-Balance(++) to ADASYN, SMOTE, and a simple balancing strategy that *reduces* the number of diseased samples to match the healthy count by random downsampling. **E3.** While E2 is to test lack of balance, E3 is concerned with biases in the composition of classification labels in the training dataset. In E3, I use the 1/5 dataset of E2-fix for training, but for the testing set, I choose samples such that 1/5, 1/2, or 4/5 are healthy (achieved by randomly removing diseased cases until the desired ratio is achieved). Thus, the last two cases have a different composition of labels between training and testing datasets.

7.3.3 Evaluation procedure.

For measuring classification accuracy, I rely on the Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC). The AUC measure is computed by exploring different cutoffs for the threshold used internally in each classification method, hence exploring the trade-off between precision and recall. AUC is the standard method used for measuring the accuracy of ML classification because it does not depend on arbitrary sensitivity/specificity trade-offs.

All of my tests are based on a cross-validation strategy, repeated several times to get a total of 20 evaluations of AUC. I report the mean and standard error of AUC across the 20 replicates. In E1 and E2-max, I use 5-fold cross-validation, repeated 4 times. In E2-fix and E3, I use 3-fold validation for the $^1/_3$ setting, 5-fold validation for $^1/_5$, and 10-fold validation, for $^1/_{10}$, each repeated enough to get 20 replicates. The augmented samples are only added to the training data, and testing is done using the held-out samples from the original datasets.

7.3.4 Method details

OTU and phylogeny. I use Deblur [8] to extract error-corrected (de-noised) sequences from each sample and take each resulting sequence as an OTU. I then use SEPP [162, 104] to insert OTUs onto a backbone phylogeny of GreenGenes [57]; removing the backbone sequences and randomly resolving the remaining polytomies gives us a binary tree on the OTUs observed in the samples. I use this tree as $\mathcal{T}$.

TADA. I implemented TADA in python using Dendropy [242] for manipulating phylogenies, biom-format [154] for processing OTU tables, scikit-learn [182] for machine learning methods, and scikit-bio for computing distances between microbiome samples. In all the analyses, I use $f(d_u) = 100\sqrt{d_u}$. The choice of the square root is arbitrary but is motivated by wanting a slower than linear reduction in variance closer to the tips (where $d_u < 1$); the constant 100 ensures the variance of Beta is not extremely high and had little impact on results in my initial tests on a different dataset. To cluster samples, I use the k-means method [15] applied to the Bray-Curtis [156] distances between samples computed from the normalized matrix $\mathbf{X}$. In all analyses, unless specified, I set $k = 5$ for TADA-SV and $k = 1, q = 5$ for TADA-TVSV-C (my initial experiments showed marginal improvements with increased k or q; see Fig. 7.3). For TADA-Balance++, I set $k = 50$ for TADA-SV and $k = 1$ and $q = 50$ for TADA-TVSV-C. For TVSV, I will explore five settings of C, the number of clusters: $1, 4, 8, 40$, and m. In order to avoid zero counts, I add the pseudocount $^{5}/_{n}$ to the count of all OTUs for all samples ($n \approx 10^4$ for IBD and $\approx 2 \times 10^4$ for BMI).

ADASYN/SMOTE. I use ADASYN and SMOTE implemented in the imbalanced-learn package (ver. 0.4.3) [129]. I use the normalized counts of OTUs (so that values in each row of $\mathbf{X}$ add up to 1) as inputs. I use $k = 5$ (default value) for the k-nearest neighbor clustering step of these methods. Both methods allow us to set the number of samples I want to generate from each class.

ML. I use two ML methods: random forests (RF) [35] and neural networks (NN), both as implemented in the scikit-learn package [182] (ver. 0.20). I use RF because of its superior performance on previous studies of microbiome (e.g., [233]). I set the number of trees for RF to 2000 and use default options otherwise. For NN, I use Multi-layer Perceptron classifier, MLPC. My MLPC had two layers with dimensions 2000 and 1000, respectively, with an early stopping rule. For the other parameters of MLPC, I used the default options. I use the normalized counts of OTUs as input features.

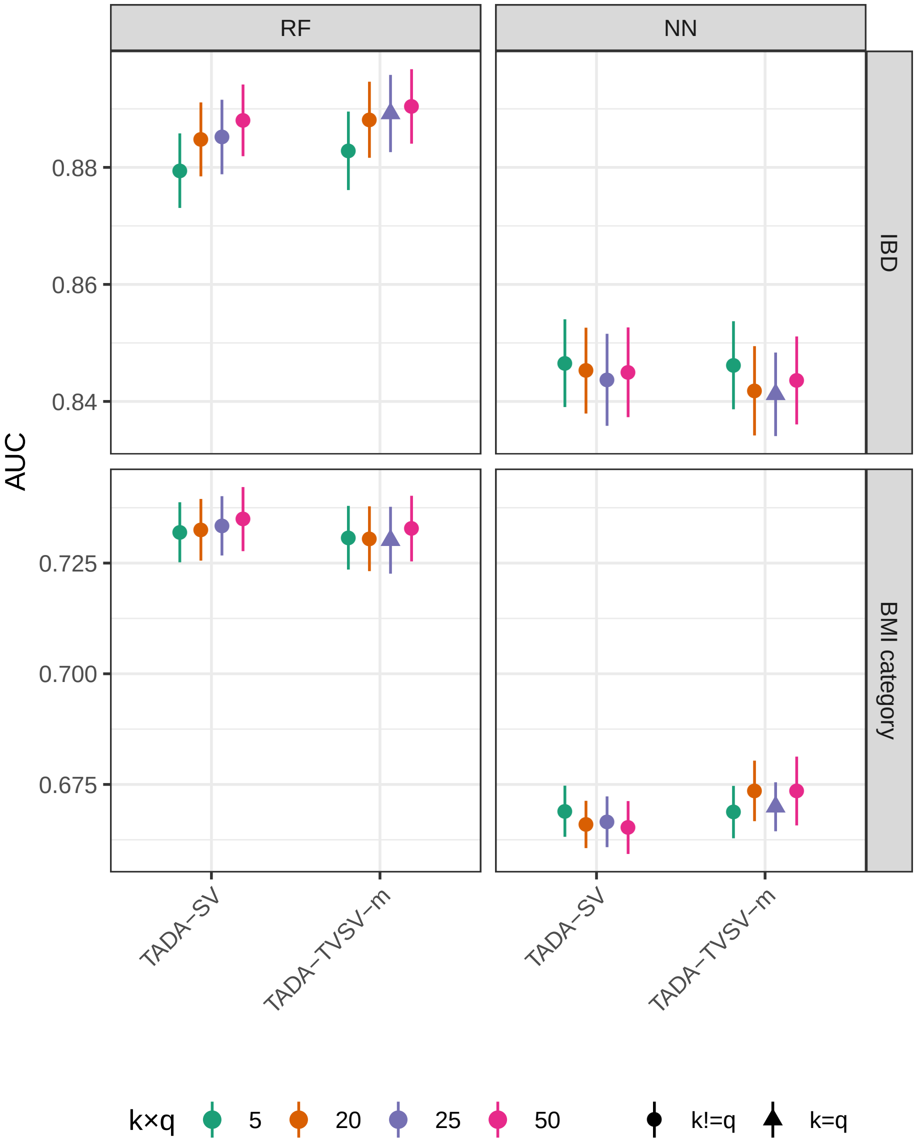


Figure 7.3: Impact of the choice of k and q. On E1, I show results with different augmentation levels. Area Under Curve (AUC) is shown for both Neural Networks(NN) and Random Forest (RF) classifiers and on both Gevers IBD dataset and AGP BMI dataset. Colors show $k \times q$ (set to 1, 5, 20, or 50). For SV, $q = 1$ in all cases. For TVSV, $k = 1$, except for the case where $k \times q = 25$, where, $k = q = 5$.

.

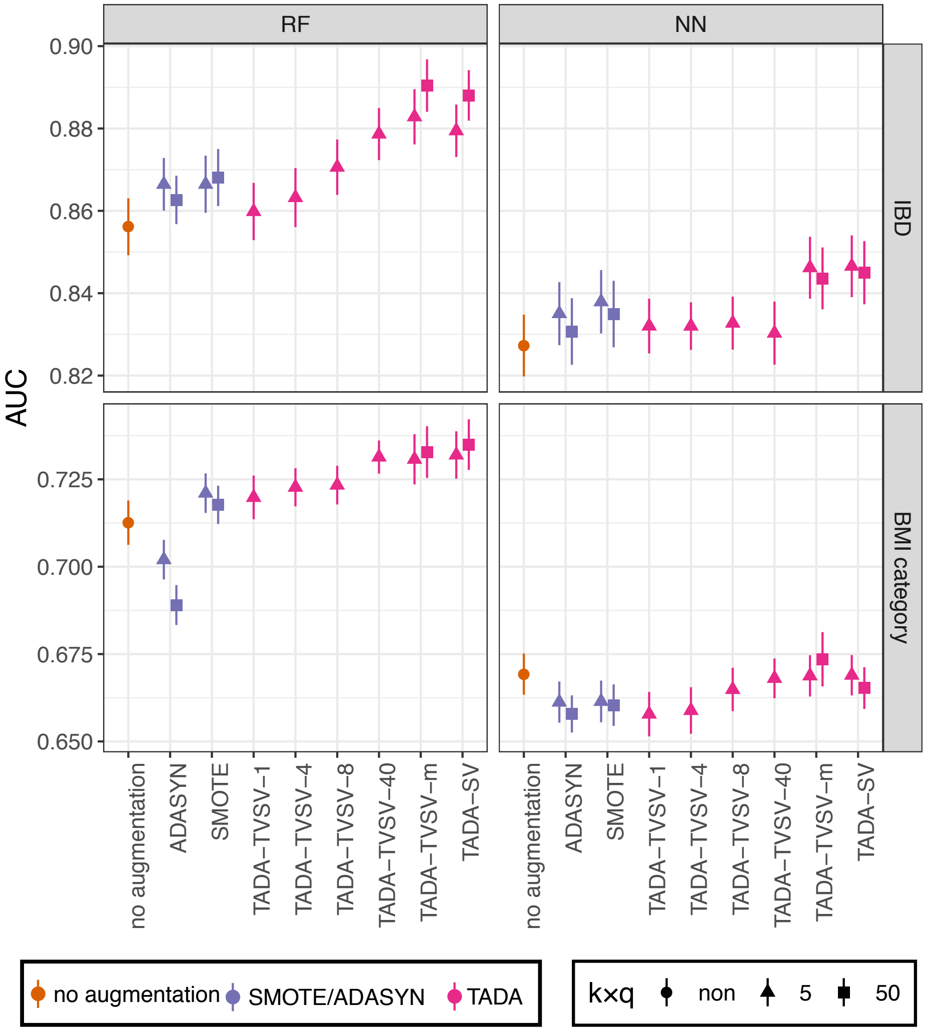


Figure 7.4: Results on E1. Area Under Curve (AUC) is shown for both Neural Networks (NN) and Random Forest (RF) classifiers and on both Gevers IBD dataset and AGP BMI dataset. I compare training on original dataset with no augmentation, SMOTE, ADASYN and using both SV and TVSV versions of TADA. For TVSV-C, I set the number of clusters, C, to 1, 4, 8, 40, or m (number of samples). I used ADASYN and SMOTE with their default settings. I show mean (dots) and standard error over 20 replicates. For TADA-SV, I show both $k = 5$ and $k = 50$, and for TADA-TVSV-m, I show both $q = 5$ and $q = 50$ with $k = 1$; see Fig. 7.3 for other q and k.

7.4 Results

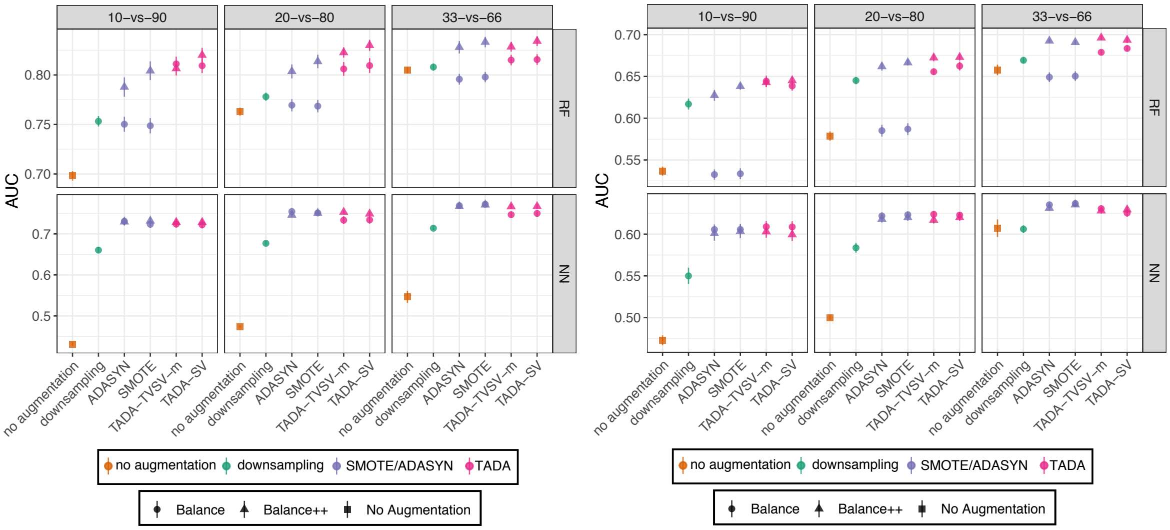
7.4.1 E1: complete datasets

I start with the E1 experiment where all the data are used (Fig. 7.4).

On the Gevers IBD dataset, the accuracy of ML methods, as measured by AUC, is reasonably high (mean AUC > 0.8, both for NN and RF) even without augmentation. Nevertheless, TADA is able to increase the mean accuracy for both NN and RF. For example, for RF, the AUC improves from 0.857 to 0.890 with TADA-TVSV-*m* ($q = 50$) and the difference is statistically significant according to a paired t-test ($p \ll 10^{-5}$). This improvement, while not large in magnitude, corresponds to a 23% reduction in the gap compared to the ideal AUC $= 1$ and therefore is substantial. In contrast, ADASYN and SMOTE result in much smaller improvements (mean AUC $<$0.87); these improvements are not statistically significant for ADASYN ($p = 0.15$) but are significant for SMOTE ($p = 0.0003$).

For BMI classification using the AGP dataset, the AUC was generally low in the absence of augmentation (mean $<$0.72 for both methods), perhaps reflecting the heterogeneous nature of the AGP dataset or the difficulty of classifying BMI into two categories based on the microbiome. Data augmentation using TADA-TVSV increases the accuracy for RF; for example, the AUC is increased to 0.73 using TADA-TVSV-*m*, and this improvement is statistically significant ($p \ll 10^{-5}$). Here, ADASYN *reduces* accuracy while SMOTE helps accuracy insignificantly ($p = 0.18$) and not as much as TADA. Unlike RF, NN is not helped by TADA-SV, and TADA-TVSV-*m* gives only a statistically insignificant improvement ($p = 0.35$). Both ADASYN and SMOTE reduce the accuracy.

Comparing different numbers of clusters (C) for TVSV-C, I observe an interesting pattern. Increasing the number of clusters improves AUC consistently, and the trend is especially apparent for RF. The highest accuracy is obtained by either TVSV-40 or TVSV-*m*



(a) IBD

(b) AGP BMI

Figure 7.5: Results on the E2-fix dataset. Training dataset is randomly subsampled to create unbalance: healthy (for IBD) and overweight (for BMI) samples constitute $^1/_{10}$ (10-vs-90), $^1/_5$ (20-vs-80), or $^1/_3$ (33-vs-66) of the samples for *both* the training and testing sets. I compare AUC on the original training set (no augmentation); the over-represented class downsampled to match the number of under-represented class (downsampling); and, augmentation using SMOTE, ADASYN, and TADA. Methods are run in two ways: TADA-Balance just adds samples to the healthy class to balance labels; TADA-Balance++ adds both healthy and unhealthy samples to make them balanced *and* to increase the total number of samples by 50X.

where a single sample (for *m*) or a handful of samples (for 40) constitute a cluster. Based on these results, I focus only on TVSV-*m* for E2 and E3. Interestingly, the accuracy of the simpler model, SV, is very close to TVSV, except perhaps on the BMI dataset with NN. Finally, increasing k or q and also using $k = q = 5$ tends to improve accuracy, albeit marginally (Fig. 7.3).

7.4.2 E2: unbalanced class labels

The power of TADA becomes evident when the classes have an unbalanced representation (Fig. 7.5). By making the representation of the two labels unbalanced, I observe that the accuracy of ML methods degrades quickly. In E2-fix, I see a sharp drop in AUC of both ML methods as the level of unbalance increases (Fig. 7.5). For example, on the IBD dataset, RF with no augmentation goes from AUC = 0.8 with $^1/_3$ healthy samples to AUC = 0.7 when $^1/_{10}$ are healthy. Similarly, on BMI, AUC goes down from 0.66 in the $^1/_3$ case to AUC = 0.54 when $^1/_{10}$ are overweight. Simply down-sampling the number of over-represented class to match the other label by random removals *increases* AUC despite training from a smaller dataset. This improved accuracy further underscores the detrimental impact of a lack of balance.

Large improvements in AUC are obtained when I use TADA to balance the representation from the two groups. For example, on the IBD dataset, the AUC of RF with TADA-SV-Balance is > 0.8 even when the original training data (i.e., before augmentation) has only $^1/_{10}$ healthy individuals. Similar levels of improvement are observed for BMI. Across both datasets, improvements in accuracy can be as large as 0.11 points for RF and 0.29 points for NN. Thus, TADA-Balance can largely erase the negative impacts of unbalance in the original training dataset. Like E1, here, TADA-SV and TVSV-*m* perform similarly.

More interestingly, using TADA-Balance++ results in additional improvements beyond TADA-Balance. For example, for IBD, the AUC in the $^1/_5$ healthy case goes from

0.81 with TADA-SV-Balance to 0.83 with TADA-SV-Balance++ with RF (statistically significant: $p = 0.00004$). The improvements of Balance++ over Balance are consistent with improvements of TADA over no augmentation observed in E1.

The two standard methods, SMOTE and ADASYN, have mixed performance. I start with RF on the IBD dataset. With the Balance version, both methods improve AUC substantially only for the $^1/_{10}$ healthy case but they fail to outperform downsampling. In the $^1/_5$ healthy case, they result in small improvements and in the $^1/_3$ healthy case they *reduce* AUC compared to no augmentation. The Balance++ versions of both methods, however, consistently improve AUC. Nevertheless, with $^1/_{10}$ or $^1/_5$ healthy, TADA-SV outperforms both methods ($p < 0.007$ in all four comparisons) whereas with $^1/_3$, TADA-SV and both methods are statistically indistinguishable ($p > 0.16$ in both comparisons). Similar patterns are observed for BMI with RF. With NN (which has much lower AUC than RF) SMOTE, ADASYN, and TADA have similar accuracy in all conditions.

The positive impact of balancing on E2-fix is not merely due to its small training set. On E2-max, which has roughly double the training set size of E2-fix, TADA continues to improve accuracy over no augmentation and other methods, especially for $^1/_{10}$ and $^1/_5$ levels of unbalance (Fig. 7.6). Compared to E2-fix, AUC is improved for all methods in E2-max, as expected due to the larger training set. Here, downsampling and SMOTE/ADASYN-Balance stop increasing accuracy for RF.

Note that before augmentation, the composition of class labels in the testing set matches that of the training set. Thus, the reductions in accuracy for unbalanced data without augmentation are not due to a biased distribution of labels in the training set. In fact, after balancing using TADA, the distribution of labels between training and testing data will not match, making the high accuracy of balanced results even more noteworthy.

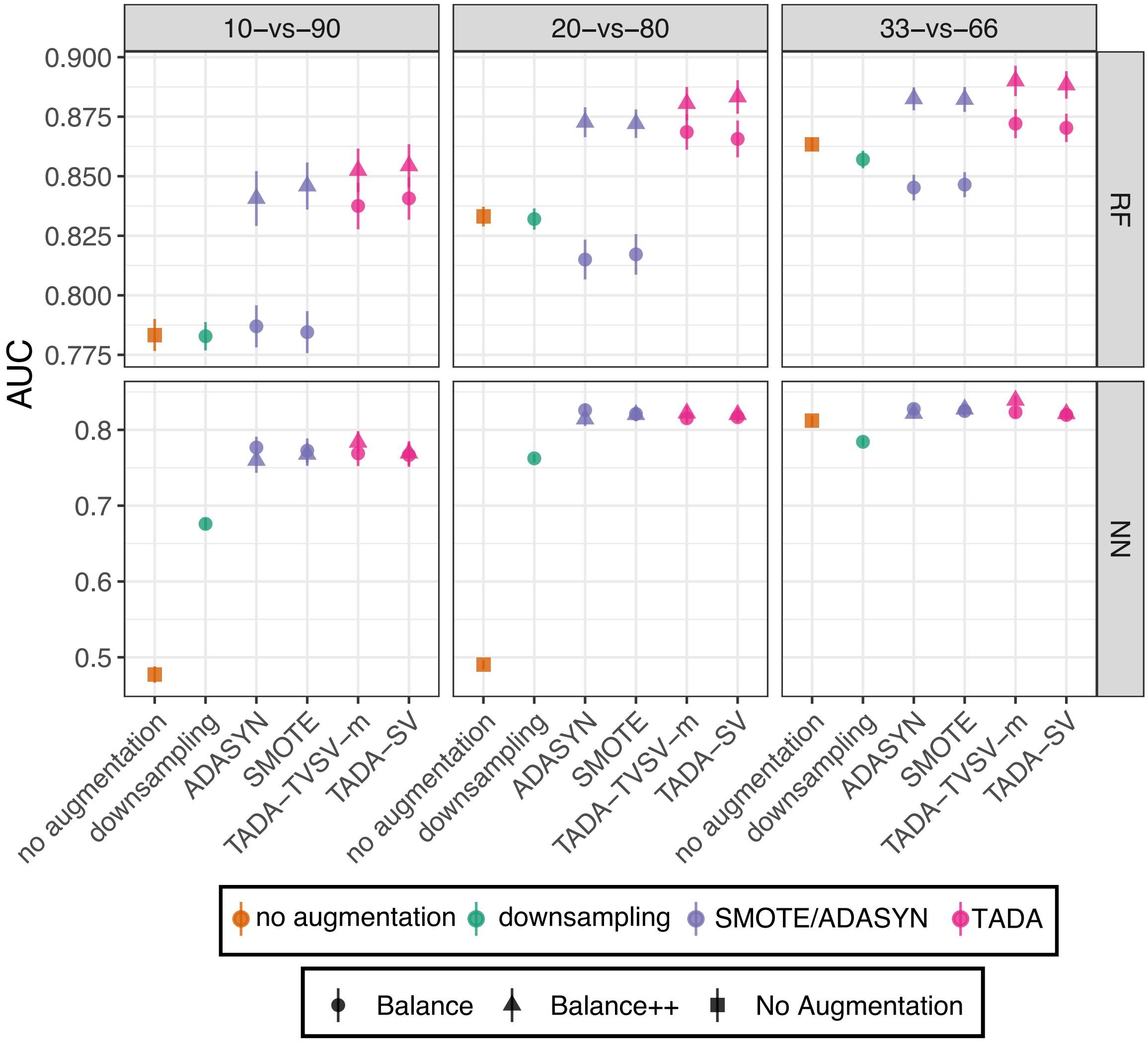


Figure 7.6: Results on the E2-max dataset. Settings similar to Figure 7.5.

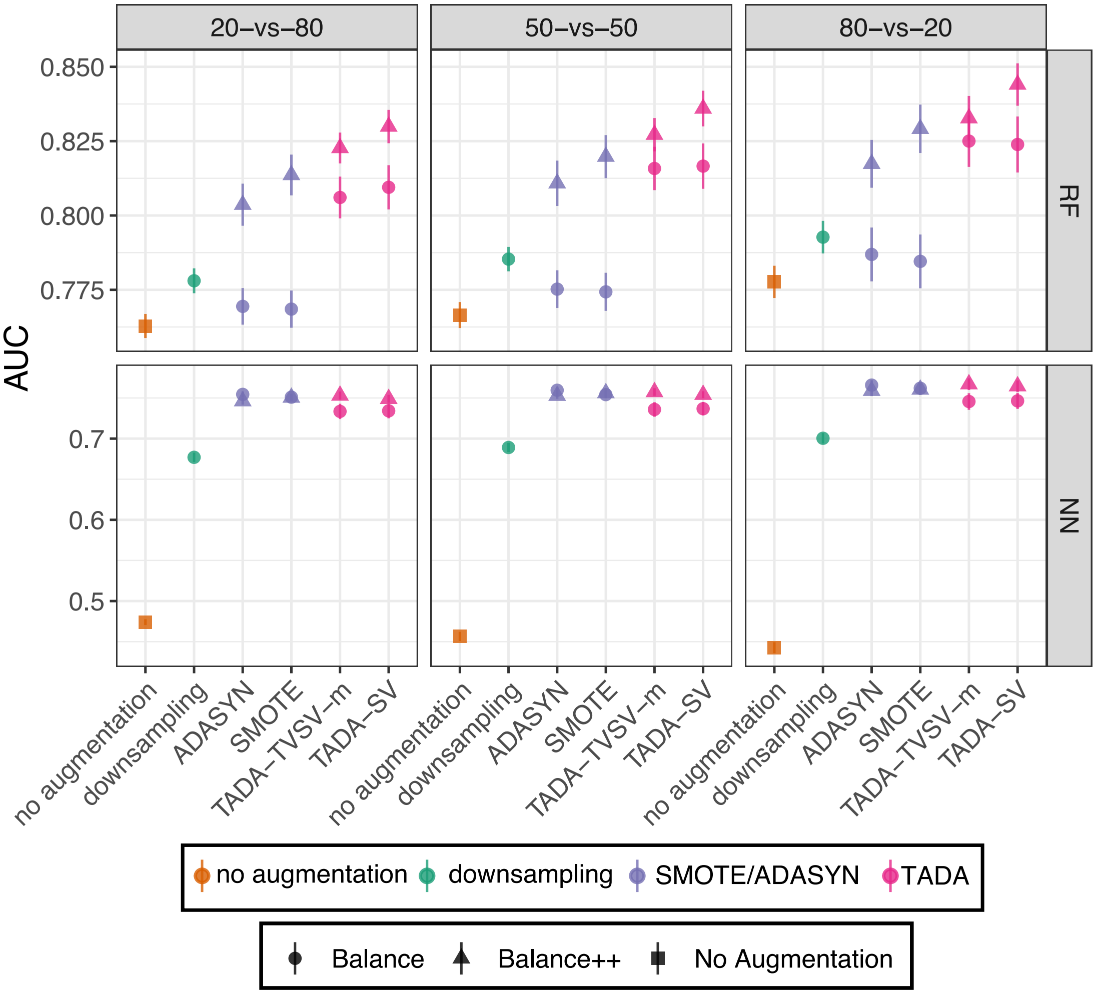


Figure 7.7: Results for E3. The training set includes 1/5 healthy out of a total of 243 samples. The testing set has 1/5 (20-vs-80), 1/2 (50-vs-50), or 4/5 (80-vs-20) of samples coming from healthy individuals on the IBD dataset. Methods labeled identically to Figure 7.5.

7.4.3 E3: biased class labels

Focusing on the IBD data, I next test the impact of not just unbalanced but also biased sampling by fixing training set to have $^1/_5$ healthy (for IBD) but changing the relative representation in testing data. Interestingly, including bias does not further reduce the accuracy in substantial ways (Fig. 7.7). However, in the biased scenario, I continue to see dramatic improvements obtained by TADA compared to no-augmentation, down-sampling, and to less extent, SMOTE and ADASYN. Thus, for unbalanced training data, augmentation can improve the accuracy regardless of whether the testing data has the same label distribution.

7.5 Discussions and conclusions

I described a new data augmentation method to generate artificial samples for augmenting the training set of ML methods for phenotype classification from microbiome samples. My method, TADA, combines the power of statistical generative models that incorporate phylogenetic knowledge with the flexibility of black-box ML methods. I tested my method for two phenotypes (IBD and BMI) and using one type of microbiome data, namely 16S. My results showed that TADA improved the classification accuracy and the improvements were dramatic when the samples were unbalanced in terms of the distribution of class labels.

I emphasize that the unbalance in training data is not a corner case; in microbiome data, unbalance is the rule, not the exception. Often, microbiome datasets gathered in clinical settings are short on control (i.e., healthy) cases, especially when compared to the larger population. My results clearly demonstrate that ML methods fail to train well on unbalanced data. While I focused on AUC, it is instructive also to examine the percentage of times a method makes the correct classification call. With $^1/_{10}$ or $^1/_5$ unbalance levels, the trained ML model is mostly useless because it classifies all testing samples as diseased, achieving artificially high levels of correct classification (Fig. 7.8) despite low AUC (Fig. 7.5); i.e., here, ML models just match a *no-skill* classifier and are, thus, grossly overfit. Balancing augmentation helps to alleviate this issue, as evident in increased AUC values. Nevertheless, balancing changes the prevalence of labels and needs to be done with care. Overall, my results provide a cautionary note on applying ML methods for unbalanced labels and are a reminder that clinical applications of ML to microbiome are fraught with dangers and can benefit from further improvements in the methodology.

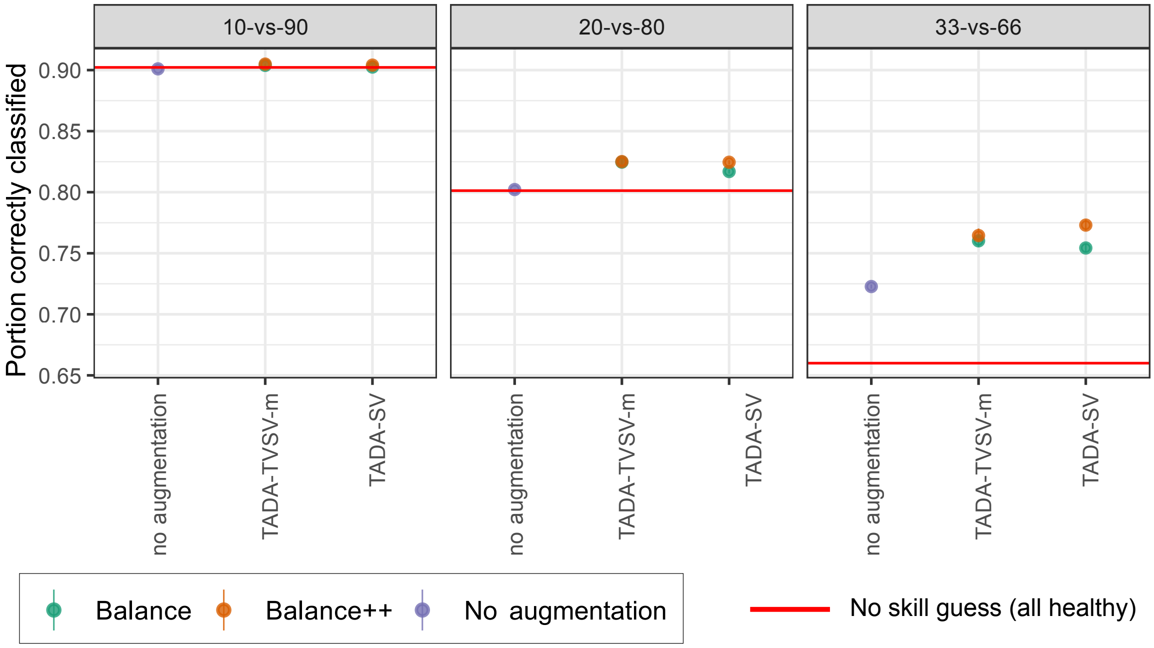


Figure 7.8: Percentage of correct classification on the E2-fix IBD dataset. Figure settings are similar to Figure 7.5 but I show percentage of correct classifications instead of AUC. Red line shows the accuracy achieved by simply guessing the healthy label each time.

My results did not show a consistent difference between SV and TVSV generative models across all dataset. The more complex model, TVSV, was slightly more accurate on the BMI dataset but did not manage to outperform SV on IBD. TVSV seeks to capture variability due to biological sources and adds more variance than SV. The failure of TVSV to provide a substantial improvement over SV only on the IBD dataset may indicate that for some datasets (perhaps more carefully curated) the biological variance is already sufficiently captured. But it could also indicate that the variance generated using my hierarchical Beta model, fails to emulate biological variance in a meaningful way. Beta is a powerful model to capture the distribution of proportions, especially when distributed around a center (or the two extremes) but biological distributions may not fit Beta. Moreover, I make conditional independence assumptions on the phylogenetic tree, which may not match the biology (e.g., due to horizontal gene transfer).

My results indicated that clustering samples and using the mixture model could reduce accuracy if the clusters are big and is neutral or only slightly beneficial when clusters are small (Fig. 7.4). Thus, sample augmentation was most effective when applied to individual samples or small clusters. It may be that with the small sample sizes that I have and large numbers of confounding factors, samples are so varied that only one or a handful of data points are available per component of the mixture model. Thus, it is possible that as the size of the training datasets increase, the mixture model starts to outperform the TVSV-*m* consistently. Thus, for existing small datasets, using TVSV-*m* is a safe choice, but in the future, as more data become available, this question needs to be revisited.

The framework I described for combining generative models and ML methods can be extended beyond the exact generative models I used. My specific generative model combines a Binomial and a Beta distribution, with one learned parameter (μ) and one parameter fixed based on the phylogeny (ν). The method I used to choose the fixed ν (inversely related to the variance) relies on the phylogenetic knowledge, incorporated as the mean divergence below each node (similar to the F_{ST} measure). I selected a particular function f, but note that my choice is without strong theoretical underpinnings. Future work should explore more principled choices, deriving the function f as a result of a dispersion process running along the branches of the phylogeny (e.g., a Poisson model). These future attempts could also explore the [18] model, which also is based on a similar Beta model and the F_{ST} measure.

A natural extension of my generative model is to let ν be learned from the data instead of using the phylogeny. In fact, I have derived the necessary parameter estimators for such a model using the method of moments (see Appendix A.2). However, using this model will double the number of parameters and will rely less on the known phylogenetic knowledge. My initial tests (Fig. 7.9) indicate this more parameter-rich model fails to perform well on my two test datasets. However, if substantially larger training sets are available in the future, this method should be revisited. Another natural extension is to use Dirichlet+Multinomial instead of Beta+Binomial to allow multifurcating trees. Finally, instead of assuming all μ_u and ν_u parameters are separate parameters, they can be considered random variables drawn from another distribution with appropriate hyperparameters.

7.5.1 Using a Beta-Binomial distribution with two parameters

Following Lemma 7, instead of using a ν_u which only depends on the phylogeny $\mathcal{T}$, I could use a model where (μ_u, ν_u) both depend only on class/cluster label y. In this model for each internal node u of the phylogeny, the probability of observing a species on the left subtree of u follows a beta distribution, (i.e. $p_u^l \sim Beta(\mu_u, \nu_u)$), and the number of observed sequences contained under the left subtree of node u follows a binomial distribution (i.e. $C^{l(u)} \sim Bin(p_u, c^u)$, where c^u is the total number of sequences under the node u). In this section, for the ease of calculation, I use the other formulation of the beta distribution, i.e. $p_u^l \sim Beta(\alpha_u, \beta_u)$, where $\mu_u = \frac{\alpha_u}{\alpha_u + \beta_u}$, and the relationship between ν_u, μ_u, and the variance of the beta distribution is given in Section 7.2.2; and hence these notations are interchangeable. Now, using method of moments, I could estimate α_u and β_u from the class-y samples as follow

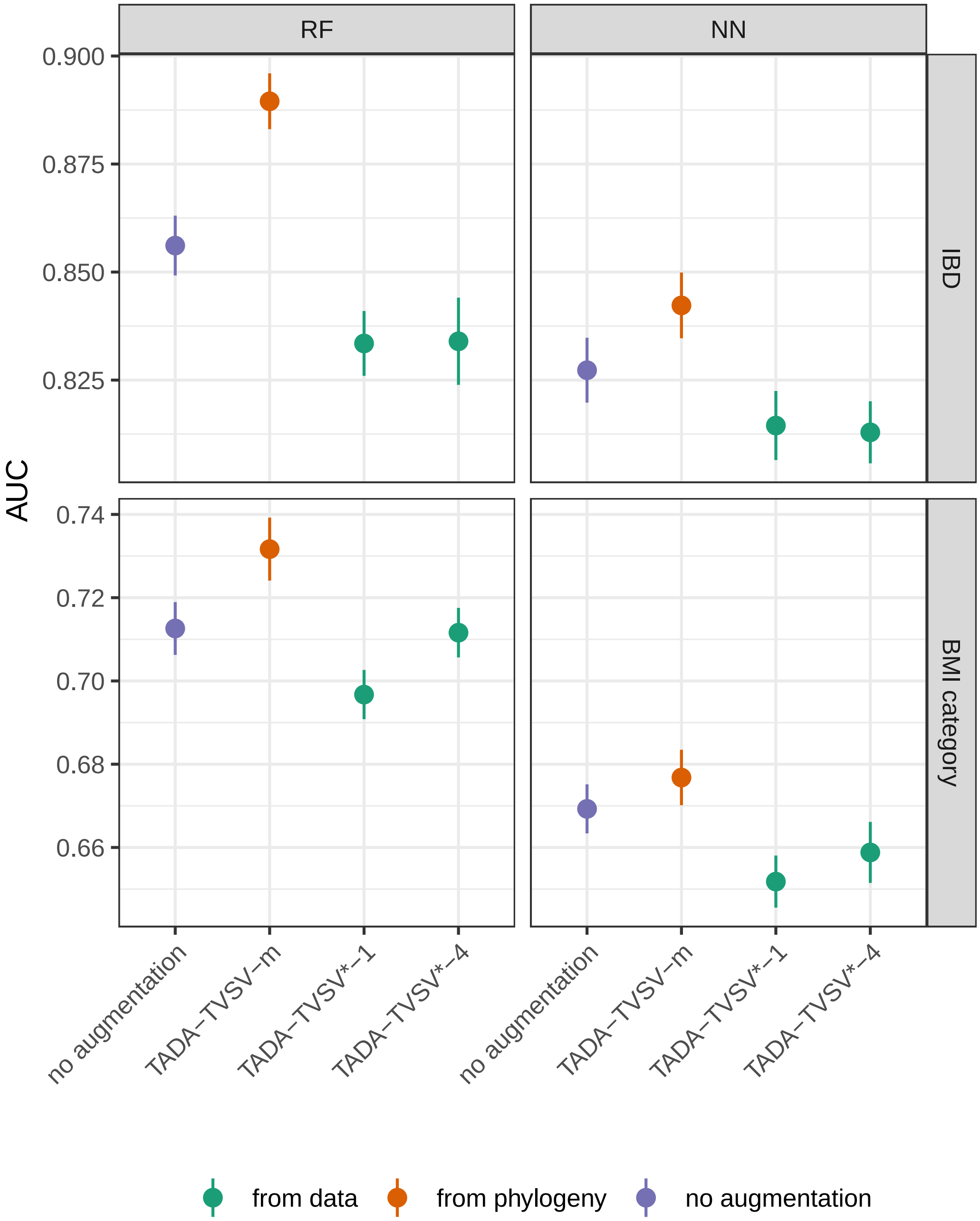


Figure 7.9: Computing ν from data versus fixing it using the phylogeny. Using the method of moments described in Section 7.5.1, I estimate both μ_u and ν_u from data (instead of fixing ν_u from phylogeny as in my main results). I call the resulting method TADA-TVSV*. Results on the E1 dataset indicate that computing ν_u from variance of the data not only fails to improve accuracy, but can even reduce it.

$$\hat{\alpha}_u^{(MOM)} = \frac{c_u^{(2)}\mathcal{M}_u - \mathcal{Q}_u c_u^{(1)}}{\frac{\mathcal{Q}_u}{\mathcal{M}_u}(c_u^{(1)})^2 - \mathcal{M}_u c_u^{(2)} + \mathcal{M}_u c_u^{(1)} - (c_u^{(1)})^2} \tag{7.4}$$

$$\hat{\beta}_u^{(MOM)} = \frac{(\frac{\mathcal{Q}_u}{\mathcal{M}_u} c_u^{(1)} - c_u^{(2)})(c_u^{(1)} - \mathcal{M}_u)}{(c_u^{(1)})^2 - \mathcal{M}_u c_u^{(1)} + \mathcal{M}_u c_u^{(2)} - \frac{\mathcal{Q}_u}{\mathcal{M}_u}(c_u^{(1)})^2} \tag{7.5}$$

$$\mathcal{M}_u = \frac{1}{w}\sum_{i=1}^{w} c_i^{l(u)} \tag{7.6}$$

$$\mathcal{Q}_u = \frac{1}{w}\sum_{i=1}^{w} (c_i^{l(u)})^2 \tag{7.7}$$

$$c_u^{(1)} = \frac{1}{w}\sum_{i=1}^{w} c_i^u \tag{7.8}$$

$$c_u^{(2)} = \frac{1}{w}\sum_{i=1}^{w} (c_i^u)^2 \tag{7.9}$$

where $\mathcal{M}_u$ and $\mathcal{Q}_u$ are the first and second moments of the observed sequences contained under the left child of the node u in $\mathcal{T}$. In a special case where the total number of observed sequences contained under the node u are all the same (i.e. $c^u = c_1^u = c_2^u = \ldots = c_w^u$), the distribution becomes the beta-binomial distribution, $c^u = c_u^{(1)}$, and $(c^u)^2 = c_u^{(2)}$. In this case, the $\alpha^{(MOM)}$ and $\beta^{(MOM)}$ equal [253]

$$\hat{\alpha}_u^{(MOM)} = \frac{c^u \mathcal{M}_u - \mathcal{Q}_u}{c^u(\frac{\mathcal{Q}_u}{\mathcal{M}_u} - \mathcal{M}_u - 1) + \mathcal{M}_u} \tag{7.10}$$

$$\hat{\beta}_u^{(MOM)} = \frac{(c^u - \mathcal{M}_u)(c^u - \frac{S}{\mathcal{M}_u})}{c^u(\frac{\mathcal{Q}_u}{\mathcal{M}_u} - \mathcal{M}_u - 1) + \mathcal{M}_u} \tag{7.11}$$

Proof. Consider the internal node u of the phylogeny $\mathcal{T}$. Based on my model, $C_i^{l(u)} \sim$

$Bin(c_i^u, p_u^l)$, where $p_u^l \sim Beta(\alpha_u, \beta_u)$

$$p_u^l \sim Beta(\alpha_u, \beta_u) \tag{7.12}$$

$$C_i^{l(u)} \sim Bin(c_i^u, p_u) \tag{7.13}$$

I will compute the expected value for the weighted average of random variables $C_i^{l(u)}$s.

$$\begin{aligned}
&\mathbb{E}[\frac{1}{w}\sum_{i=1}^{w} C_i^{l(u)}] = \frac{1}{w}\sum_{i=1}^{w}\mathbb{E}[C^{l(u)}] = \frac{1}{w}\sum_{i=1}^{w}\mathbb{E}[\mathbb{E}[C^{l(u)}|p_u^l]] = \\
&\frac{1}{w}\sum_{i=1}^{w} c_i^u\mathbb{E}[p_u^l] = \frac{\alpha_u}{\alpha_u+\beta_u}\frac{1}{w}\sum_{i=1}^{w} c^u = \frac{\alpha_u}{\alpha_u+\beta_u}c_u^{(1)}
\end{aligned}$$

I can name the empirical mean of $C^{l(u)}$s', as $\mathcal{M}_u = \frac{1}{w}\sum_{i=1}^{w} c_i^{l(u)}$. Next I write the expected value of $(C_i^{l(u)})^2$s'

$$\begin{aligned}
&\mathbb{E}[\frac{1}{w}\sum_{i=1}^{w}(C_i^{l(u)})^2] = \frac{1}{w}\sum_{i=1}^{w}\mathbb{E}[(C_i^{l(u)})^2] = \frac{1}{w}\sum_{i=1}^{w}\mathbb{E}[\mathbb{E}[(C_i^{l(u)})^2|p_u^l]] \\
&\mathbb{E}[(C_i^{l(u)})^2|p_u^l] = c_i^u p_u^l(1+(c_i^u-1)p_u^l) \\
&\mathbb{E}[\mathbb{E}[(C_i^{l(u)})^2|p_u^l]] = c_i^u\mathbb{E}[p_u^l] + c_i^u(c_i^u-1)\mathbb{E}[(p_u^l)^2] = \\
&c_i^u\frac{\alpha_u}{\alpha_u+\beta_u} + c_i^u(c_i^u-1)(\frac{\alpha_u\beta_u}{(\alpha_u+\beta_u)^2(\alpha_u+\beta_u+1)} + \frac{\alpha_u^2}{(\alpha_u+\beta_u)^2}) = \\
&c_i^u\alpha_u(\frac{c_i^u(\alpha_u+1)+\beta_u}{(\alpha_u+\beta_u)(\alpha_u+\beta_u+1)})
\end{aligned}$$

hence

$$\mathbb{E}[\frac{1}{w}\sum_{i=1}^{w}(C_i^{l(u)})^2] = \sum_{i=1}^{w}\frac{1}{w}(c_i^u\alpha_u(\frac{c_i^u(\alpha_u+1)+\beta_u}{(\alpha_u+\beta_u)(\alpha_u+\beta_u+1)})) = \tag{7.14}$$

$$\frac{\alpha_u}{(\alpha_u+\beta_u)(\alpha_u+\beta_u+1)}((\alpha_u+1)\sum_{i=1}^{w}\frac{(c_i^u)^2}{w} + \beta_u\sum_{i=1}^{w}\frac{c_i^u}{w}) = \tag{7.15}$$

$$\frac{\alpha_u}{(\alpha_u+\beta_u)(\alpha_u+\beta_u+1)}((\alpha_u+1)c_u^{(2)} + \beta_u c_u^{(1)}) \tag{7.16}$$

where I call the empirical value for the $\mathbb{E}[\frac{1}{w}\sum_{i=1}^{w} C_i^{l(u)}]$ as $\mathcal{Q}_u = \sum_{i=1}^{w}\frac{1}{w}(c_i^{l(u)})^2$. Using the method of moments and using the following equations I could compute the estimates for α and β

$$\mathcal{M}_u = \frac{\alpha_u}{\alpha_u+\beta_u} c_u^{(1)} \tag{7.17}$$

$$\mathcal{Q}_u = \frac{\alpha_u}{(\alpha_u+\beta_u)(\alpha_u+\beta_u+1)}((\alpha_u+1)c_u^{(2)}+\beta_u c_u^{(1)}) = \frac{\mathcal{M}_u c_u^{(2)}}{c_u^{(1)}}+ \tag{7.18}$$

$$(c_u^{(1)}-c_u^{(2)})\frac{\mathcal{M}_u}{c_u^{(1)}}\frac{\beta_u}{\beta_u+\alpha_u+1} \tag{7.19}$$

$$c_u^{(1)} = \frac{1}{w}\sum_{i=1}^{w} c_i^u \tag{7.20}$$

$$c_u^{(2)} = \frac{1}{w}\sum_{i=1}^{w} (c_i^u)^2 \tag{7.21}$$

□

In order to evaluate the performance of these estimators, I used simulations, where μ's and counts (c^u's) are generated from known beta and binomial distributions respectively (Figure S1). Figure S4 shows results of these estimators on the IBD dataset.

I observed that RF had somewhat higher accuracy than NN in my experiments. This observation is in line with some previous studies (e.g., [233]), which have demonstrated similar results. However, I note that with augmentation, NN comes much closer to the accuracy of RF. I also note that I have not fine tuned the NN models. Thus, it is possible that NN, perhaps in the form of smaller networks or conversely deeper networks along with regularization techniques could outperform RF. In particular, deep learning requires large training samples. It is conceivable that deep learning methods paired with augmented data can in the future outperform ensemble methods such as NN in the future.

Other steps of TADA could also be changed. For example, for the phylogenetic inference, instead of placement on a common backbone tree, a *de novo* inference may be

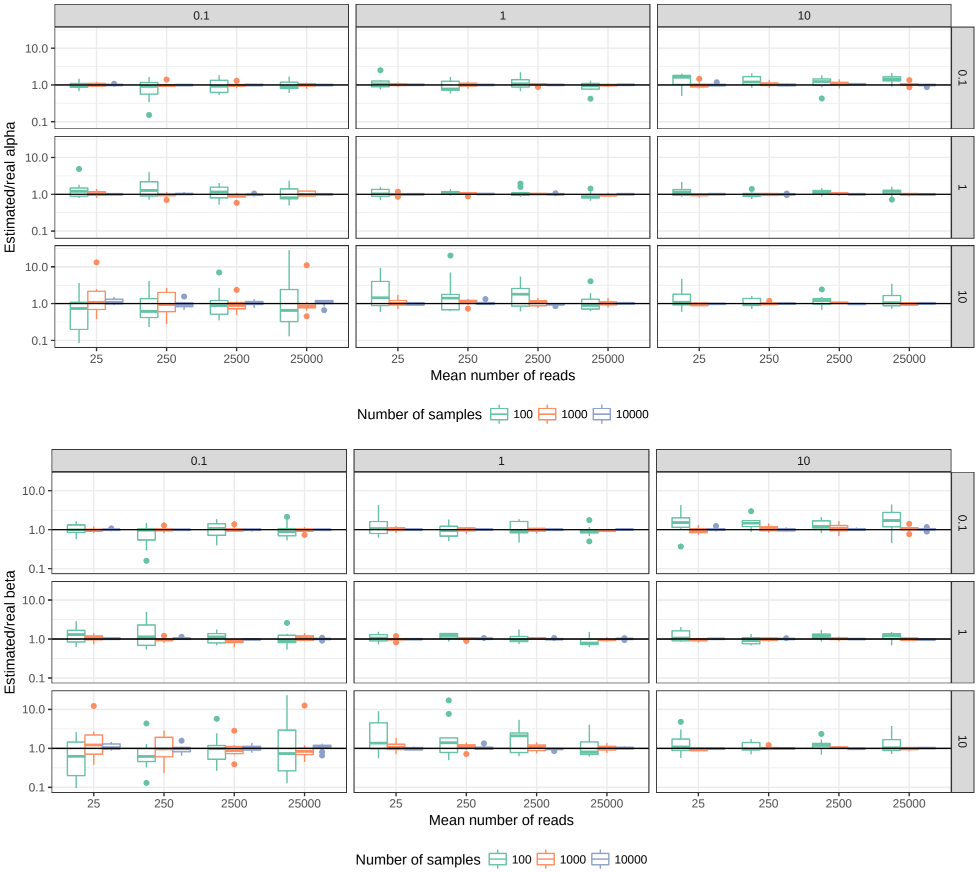


Figure 7.10: The α estimation error using the method of moments in simulations. 10,000 points are drawn from the hierarchical model and accuracy of the method of moments estimator is shown. The x-axis shows the average number of reads, the y-axis shows the ratio between estimated and the real α (top) and β (bottom). Each row corresponds to the true α values and each column corresponds to the true β values.

feasible using scalable phylogenetic inference methods. Clustering of samples can also be done using more complex methods designed for microbiome, such as phylogeny-based methods like weighted/unweighted Unifrac distances [145, 144] and compositional methods like Aitchison's distance [3, 4]. Finally, as features for the ML training, I used OTUs as obtained using the Deblur algorithm (i.e., de-noised sequences). However, extracting features can also follow more complex methods, perhaps using those that include the phylogenetic knowledge (e.g., [5, 168]).

My studies show potential for improving the generalization of ML methods. I tested only two datasets, each with only two categories. Future work should explore applications of TADA to more phenotypes, including multi-labeled ones. Also, nothing in the method limits it to gut or human microbiome; the same method should be explored on other types of environments. Future experiments should also explore training models on a dataset and testing on a separate dataset produced by a different lab; perhaps augmentation can also help reduce batch effects, which are notoriously difficult to deal with in microbiome modeling. Finally, I focused on 16S profiling. However, phylogenetic placement methods for shotgun metagenomic samples also exist (e.g., TIPP [173]); future work should explore the application of TADA to metagenomic data.

7.6 Acknowledgements

Chapter 7, in full, contains material from SAYYARI, E., KAWAS, B., AND MIRARAB, S., TADA: Phylogenetic Augmentation of Microbiome Samples Enhances Phenotype Classification. Bioinformatics (ISMB Species Issue) that is accepted at ISMB 2019 and will appear in Bioinformatics journal. I was the primary investigator and author of this paper.

Appendix A

Anchored distances for quartet-based estimation of phylogenetic trees and applications to coalescent-based analyses

A.1 Commands and version numbers

We used ASTRAL version 4.7.8 to find the species trees from gene trees:

```
java -Xmx2000M -jar astral.4.7.8.jar -i [GENE TREES] -o
[OUTPUT SPECIES TREE]
```

The ASTRID (NJst) results were produced using the following command:

```
python ASTRID.py -i [GENE TREES] -m [fastme2] -o
[OUTPUT SPECIES TREE] -c  [CACHE]
```

CACHE is the distance matrix produced by ASTRID.

For DISTIQUE-allpairs we used the following command:

```
python distique.py -a [mean] -g [GENE TREES] -m [prod]  -o
```

```
[OUTPUT DIR] -t [1]
```

For DISTIQUE-allpairs-max we used the following command:

```
python distique.py -a [mean] -g [GENE TREES] -m [min]  -o
[OUTPUT DIR] -t [2]
```

Here the flag *a* specifies which averaging method to use the partial quartet tables from complete quartet tables around each polytomy.

For tree-sum we used the following command:

```
python distique.py -g [GENE TREES] -o [OUTPUT DIR] -n
[# rounds of anchoring]  -t[4]
```

For distance-sum we used the following command:

```
python distique.py -g [GENE TREES] -o [OUTPUT DIR] -n
[# rounds of anchoring]  -t[3]
```

Finally the flag *t* determines which method to use for inferring.

For comparison and computing false negative missing branches we used the command available at `https://github.com/smirarab/global/tree/master/src/shell`:

```
compareTrees.missingBranch [SPECIES TREE]
[ESTIMATED SPECIES TREE]
```

Appendix B

Supplementary Material for “Fast coalescent-based computation of local branch support from quartet frequencies”

B.1 Commands and version numbers

B.1.1 ASTRAL

We used ASTRAL version 4.9.1 available at https://github.com/smirarab/ASTRAL/tree/posteval for scoring. We also used ASTRAL version 4.9.8 for computing the branch lengths of the trees. To have posterior probabilities of branches of main species tree and 2 other alternatives we used:

```
java –Xmx2000M −jar astral.4.9.1.jar −i [GENE TREES] −q
[SPECIES TREE] −t 4
```

To compute the branch lengths of main species tree we used the MAP estimate with the command:

```
java –Xmx2000M −jar astral.4.9.8.jar −i [GENE TREES] −q
[SPECIES TREE] −t 2
```

Users can find the most updated code available at https://github.com/smirarab/ASTRAL/. To score and compute the branch lengths, and local posterior probabilities of inferred species tree one can use:

```
java –Xmx2000M −jar astral.4.10.0 −i [GENE TREES] −q
[SPECIES TREE] −t 3
```

B.1.2 MP-EST

MP-EST version 1.5 was used for estimating branch lengths on a fixed topology. We used a custom shell script to run MP-EST 2 times with different random seed numbers and take the tree with the highest likelihood. The shell script is available at https://github.com/smirarab/global/tree/master/src/shell.

Appendix C

Supplementary Material for “DiscoVista: interpretable visualizations of gene tree discordance”

C.1 Structure of parameter files

splits definitions: The splits definition file has 7 columns separated with tabs. In the first column split names are listed, in the second column the species or other splits that define this split are listed separated with “+” or “-” signs. “+” signs are used to add splits and species, and “-” signs to subtract species or splits. In the third column (you might leave this column blank) a name can be defined that is used to specify the part of the tree that the splits belongs to, e.g. 1-Base, or Base. Forth column defines splits components where in the absence of any of them splits is considered as missing. The fifth column indicates whether the split should be shown in the species tree (gene trees) analysis. The sixth column is used to define components from the other side of the split, where in the absence of any of them the split is considered as missing. Note that if you leave this column blank, “All” minus this split will be considered as the other side component. The last column is used to add any comments to this file. Also note that, a split with the name "All" should be defined which consists of all the species names in your analysis.

Annotation file: In this file, there should be a row for each species. The first column of the annotation file specifies the name of species, and the second column defines the split that this species belongs to. Columns are separated by tab.

C.2 Installation, Commands, and Usage

In order to install DiscoVista from source please refer to `https://github.com/esayyari/DiscoVista`. DiscoVista relies on different R, python , some C packages, as well as ASTRAL. In order to make the installation easier we also provide a docker image available at `https://hub.docker.com/r/esayyari/discovista/`. In order to use this image, after installing docker, it is sufficient to download it using the following command:

```
docker pull esayyari/discovista
```

Now, you would run different analyses using the following command:

```
docker run -v <absolute path to data folder>:/data
esayyari/discovista  discoVista.py [OPTIONS]
```

Species discordance analysis: For this analysis you need the path to the species trees folder, the clade definitions, the output folder, and a threshold. For more information regarding the structure of the species tree folder please refer to `https://github.com/esayyari/DiscoVista`. Then you would use the following command using docker:

```
docker run -v <absolute path to data folder>:/data esayyari/discovista
discoVista.py  -m 0  -p <path to species trees folder> -t <threshold>
-c <path to split definition>  -o <output folder> -y
<model condition ordering file> -w <splits ordering file>
```

Discordance analysis on gene trees: For this analysis you need the path to the gene trees folder, the clade definitions, the output folder, and a threshold. For more information regarding the structure of the gene trees folder please refer to `https://github.com/esayyari/DiscoVista`. Then you would use the following command with docker:

```
 docker run -v <absolute path to data folder>:/data esayyari/discovista
 discoVista.py -m 1 -p <path to gene trees folder> -t <threshold> -c
 <path to split definition> -o <output folder> -w <splits ordering file>
```

GC content analysis: For this analysis you need the path to the coding alignments (in FASTA format) with the structure described in more details at `https://github.com/esayyari/DiscoVista`, and the output folder. You would use the following command in docker:

```
 docker run -v <absolute path to data folder>:/data esayyari/discovista
 discoVista.py -m 2 -p <path to alignments folder>  -o <output folder>
```

Occupancy analysis: For this analysis you need the path to the alignments (in FASTA format) with the structure described in more details at `https://github.com/esayyari/DiscoVista`, and the output folder. Then you can use this command with the docker:

```
docker run -v <absolute path to data folder>:/data esayyari/discovista
 discoVista.py -p $path -m 3
 -a <path to annotation file> -o <output folder>
```

Relative frequencey analysis: For this analysis you need the path to the gene trees and species tree folder (structure described on the Github), annotation file, outgroup clade name, output folder. Using docker you would do this analysis with the following command:

```
docker run -v <absolute path to data folder>:/data esayyari/discovista
discoVista.py -p $path -m 5
-a <path to annotation file> -o <output folder>
```

If desired, the cartoon tree can be shown rooted, requiring the name of the outgroup as input (e.g. Base or Outgroup).

```
docker run -v <absolute path to data folder>:/data esayyari/discovista
discoVista.py  -p $path -m 5 -a <path to annotation file> -o
<output folder> -g <outgroup clade name>
```

Bibliography

[1] AAGAARD, K., RIEHLE, K., MA, J., SEGATA, N., MISTRETTA, T.-A., COARFA, C., RAZA, S., ROSENBAUM, S., DEN VEYVER, I., MILOSAVLJEVIC, A., GEVERS, D., HUTTENHOWER, C., PETROSINO, J., AND VERSALOVIC, J. A Metagenomic Approach to Characterization of the Vaginal Microbiome Signature in Pregnancy. *PLoS ONE 7*, 6 (2012), e36466.

[2] ABABNEH, F., JERMIIN, L. S., MA, C., AND ROBINSON, J. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics 22*, 10 (2006), 1225–1231.

[3] AITCHISON, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological) 44*, 2 (1982), 139–177.

[4] AITCHISON, J., BARCELÓ-VIDAL, C., MARTÍN-FERNÁNDEZ, J. A., AND PAWLOWSKY-GLAHN, V. Logratio Analysis and Compositional Distance. *Mathematical Geology 32*, 3 (2000), 271–275.

[5] ALBANESE, D., DE FILIPPO, C., CAVALIERI, D., AND DONATI, C. Explaining Diversity in Metagenomic Datasets by Phylogenetic-Based Feature Weighting. *PLOS Computational Biology 11*, 3 (2015), 1–18.

[6] ALLMAN, E. S., DEGNAN, J. H., AND RHODES, J. A. Determining species tree topologies from clade probabilities under the coalescent. *Journal of Theoretical Biology 289*, 1 (2011), 96–106.

[7] ALLMAN, E. S., DEGNAN, J. H., AND RHODES, J. A. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *J. Math. Biol. 62* (2011), 833–862.

[8] AMIR, A., MCDONALD, D., NAVAS-MOLINA, J. A., KOPYLOVA, E., MORTON, J. T., ZECH XU, Z., KIGHTLEY, E. P., THOMPSON, L. R., HYDE, E. R., GONZALEZ, A., AND KNIGHT, R. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems 2*, 2 (2017), 00191–16.

[9] ANDERSON, D. R., BURNHAM, K. P., AND THOMPSON, W. L. Null Hypothesis Testing: Problems, Prevalence, and an Alternative. *The Journal of Wildlife Management 64*, 4 (2000), 912.

[10] ANDRADE, S. C. S., NOVO, M., KAWAUCHI, G. Y., WORSAAE, K., PLEIJEL, F., GIRIBET, G., AND ROUSE, G. W. Articulating "archiannelids": Phylogenomics and annelid relationships, with emphasis on meiofaunal taxa. *Molecular Biology and Evolution* (2015).

[11] ANÉ, C., LARGET, B. R., BAUM, D. A., SMITH, S. D., AND ROKAS, A. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution 24*, 2 (2007), 412–426.

[12] ANISIMOVA, M., GASCUEL, O., AND SULLIVAN, J. Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Systematic Biology 55*, 4 (2006), 539–552.

[13] ANISIMOVA, M., GIL, M., DUFAYARD, J. F., DESSIMOZ, C., AND GASCUEL, O. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic Biology* (2011).

[14] ARNTZEN, J. W., THEMUDO, G. E., AND WIELSTRA, B. The phylogeny of crested newts (Triturus cristatus superspecies) nuclear and mitochondrial genetic characters suggest a hard polytomy, in line with the paleogeography of the centre of origin. *Contributions to Zoology 76*, 4 (2007).

[15] ARTHUR, D., AND VASSILVITSKII, S. K-Means++: the Advantages of Careful Seeding. In *Proc ACM-SIAM symposium on discrete algorithms.* (New Orleans, Louisiana, 2007), p. 1027–1035.

[16] ATTESON, K. The Performance of Neighbor-Joining Methods of Phylogenetic Reconstruction. *Algorithmica 25*, 2-3 (1999), 251–278.

[17] AVNI, E., COHEN, R., AND SNIR, S. Weighted Quartets Phylogenetics. *Systematic biology 64*, 2 (2015), 233–242.

[18] BALDING, D. J., AND NICHOLS, R. A. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica 96*, 1-2 (1995), 3–12.

[19] BAPTESTE, E., VAN IERSEL, L., JANKE, A., KELCHNER, S., KELK, S., MCINERNEY, J. O., MORRISON, D. A., NAKHLEH, L., STEEL, M., STOUGIE, L., AND WHITFIELD, J. Networks: Expanding evolutionary thinking, 2013.

[20] BAYZID, M. S., MIRARAB, S., BOUSSAU, B., AND WARNOW, T. Weighted Statistical Binning: Enabling Statistically Consistent Genome-Scale Phylogenetic Analyses. *PLoS ONE 10*, 6 (2015), e0129183.

[21] BAYZID, M. S., AND WARNOW, T. Estimating optimal species trees from incomplete gene trees under deep coalescence. *Journal of Computational Biology 19*, 6 (2012), 591–605.

[22] BECK, D., AND FOSTER, J. A. Machine Learning Techniques Accurately Classify Microbial Communities by Bacterial Vaginosis Characteristics. *PLoS ONE 9*, 2 (2014), e87830.

[23] BENJAMINI, Y., AND HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. 57* (1995), 289–300.

[24] BEUTEL, R. G., FRIEDRICH, F., HÖRNSCHEMEYER, T., POHL, H., HÜNEFELD, F., BECKMANN, F., MEIER, R., MISOF, B., WHITING, M. F., AND VILHELMSEN, L. Morphological and molecular evidence converge upon a robust phylogeny of the megadiverse Holometabola. *Cladistics 27*, 4 (8 2011), 341–355.

[25] BEUTEL, R. G., FRIEDRICH, F., YANG, X.-K., AND GE, S.-Q. *Insect morphology and phylogeny: a textbook for students of entomology*. Walter de Gruyter, 2014.

[26] BEUTEL, R. G., KRISTENSEN, N. P., AND POHL, H. Resolving insect phylogeny: The significance of cephalic structures of the Nannomecoptera in understanding endopterygote relationships. *Arthropod Structure and Development 38*, 5 (2009), 427–460.

[27] BININDA-EMONDS, O. R. P., Ed. *Phylogenetic Supertrees: combining information to reveal the tree of life*, vol. 4. Kluwer Academic Publishers, 2004.

[28] BITSCH, C., AND BITSCH, J. Phylogenetic relationships of basal hexapods among the mandibulate arthropods: A cladistic analysis based on comparative morphological characters. *Zoologica Scripta 33*, 6 (11 2004), 511–550.

[29] BOETTIGER, C., AND CARL. An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review 49*, 1 (1 2015), 71–79.

[30] BONETTA, L. Whole-Genome sequencing breaks the cost barrier. *Cell 141*, 6 (2010), 917–919.

[31] BOUCKAERT, R. R. DensiTree: Making sense of sets of phylogenetic trees. *Bioinformatics 26*, 10 (2010), 1372–1373.

[32] BOUSSAU, B., AND GOUY, M. Efficient likelihood computations with nonreversible models of evolution. *Systematic Biology 55*, 5 (2006), 756–768.

[33] BOUSSAU, B., WALTON, Z., DELGADO, J. A., COLLANTES, F., BEANI, L., STEWART, I. J., CAMERON, S. A., WHITFIELD, J. B., JOHNSTON, J. S., HOLLAND, P. W. H., BACHTROG, D., KATHIRITHAMBY, J., AND HUELSENBECK, J. P. Strepsiptera, phylogenomics and the long branch attraction problem. *PLoS ONE 9*, 10 (10 2014), e107709.

[34] BRAUN, E. L., AND KIMBALL, R. T. Polytomies, the power of phylogenetic inference, and the stochastic nature of molecular evolution: a comment on Walsh et al.(1999). *Evolution 55*, 6 (2001), 1261–1263.

[35] BREIMAN, L. Random Forrests. *Machine learning 45* (2001), 5–32.

[36] BRODAL, G. S., FAGERBERG, R., ÖSTLIN, A., PEDERSEN, C. N. S., AND RAO, S. S. Computing Refined Buneman Trees in Cubic Time. *Lecture Notes in Computer Science 2812* (2003), 259–270.

[37] BRUNO, W. J., SOCCI, N. D., AND HALPERN, A. L. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution 17*, 1 (2000), 189–197.

[38] BRYANT, D., BOUCKAERT, R., FELSENSTEIN, J., ROSENBERG, N. A., AND ROY-CHOUDHURY, A. Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Molecular Biology and Evolution 29*, 8 (8 2012), 1917–1932.

[39] BRYANT, D., AND STEEL, M. Constructing Optimal Trees from Quartets. *Journal of Algorithms 38* (2001), 237–259.

[40] BUNEMAN, P. A note on the metric properties of trees. *Journal of Combinatorial Theory, Series B 17*, 1 (1974), 48–50.

[41] CALLAHAN, B. J., MCMURDIE, P. J., ROSEN, M. J., HAN, A. W., JOHNSON, A. J. A., AND HOLMES, S. P. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods 13* (5 2016), 581–583.

[42] CAMERON, S. L., BARKER, S. C., AND WHITING, M. F. Mitochondrial genomics and the new insect order Mantophasmatodea. *Molecular phylogenetics and evolution 38*, 1 (2006), 274–279.

[43] CANNON, J. T., VELLUTINI, B. C., SMITH, J., RONQUIST, F., JONDELIUS, U., AND HEJNOL, A. Xenacoelomorpha is the sister group to Nephrozoa. *Nature 530*, 7588 (2016), 89–93.

[44] CAPORASO, J. G., LAUBER, C. L., COSTELLO, E. K., BERG-LYONS, D., GONZALEZ, A., STOMBAUGH, J., KNIGHTS, D., GAJER, P., RAVEL, J., FIERER, N., GORDON, J. I., AND KNIGHT, R. Moving pictures of the human microbiome. *Genome Biology 12*, 5 (2011), R50.

[45] CHAWLA, N. V. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Springer US, Boston, MA, 2010, pp. 875–886.

[46] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research 16* (2002), 321–357.

[47] Chifman, J., and Kubatko, L. S. Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics 30*, 23 (8 2014), 3317–3324.

[48] Chojnowski, J. L., Kimball, R. T., and Braun, E. L. Introns outperform exons in analyses of basal avian phylogeny using clathrin heavy chain genes. *Gene 410*, 1 (2008), 89–96.

[49] Criscuolo, A., and Gascuel, O. Fast NJ-like algorithms to deal with incomplete distance matrices. *BMC bioinformatics 9*, 1 (2008), 166.

[50] Darwin, C. *The origin of species by means of natural selection.* J. Murray, 1872.

[51] Dasarathy, G., Nowak, R., and Roch, S. Data requirement for phylogenetic inference from multiple loci: a new distance method. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 12*, 2 (2015), 422–432.

[52] Dave, M., Higgins, P. D., Middha, S., and Rioux, K. P. The human gut microbiome: current knowledge, challenges, and future directions. *Translational Research 160*, 4 (2012), 246–257.

[53] Degnan, J. H. Anomalous unrooted gene trees. *Systematic Biology 62* (2013), 574–590.

[54] Degnan, J. H., DeGiorgio, M., Bryant, D., and Rosenberg, N. A. Properties of Consensus Methods for Inferring Species Trees from Gene Trees. *Systematic Biology 58*, 1 (2009), 35–54.

[55] Degnan, J. H., and Rosenberg, N. A. Discordance of species trees with their most likely gene trees. *PLoS genetics 2*, 5 (2006), e68.

[56] Degnan, J. H., and Rosenberg, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution 24*, 6 (2009), 332–340.

[57] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl. Environ. Microbiol. 72*, 7 (2006), 5069–5072.

[58] Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics 26*, 19 (10 2010), 2460–2461.

[59] EDGAR, R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* (2016), 081257.

[60] EDWARDS, S. V. Is a new and general theory of molecular systematics emerging? *Evolution 63*, 1 (2009), 1–19.

[61] EDWARDS, S. V., XI, Z., JANKE, A., FAIRCLOTH, B. C., MCCORMACK, J. E., GLENN, T. C., ZHONG, B., WU, S., LEMMON, E. M., LEMMON, A. R., LEACHÉ, A. D., LIU, L., AND DAVIS, C. C. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution 94* (2016), 447–462.

[62] ELIAS, I., AND LAGERGREN, J. Fast neighbor joining. *Theoretical Computer Science 410*, 21-23 (2009), 1993–2000.

[63] ENGEL, M. S., AND GRIMALDI, D. A. New light shed on the oldest insect. *Nature 427*, 6975 (2 2004), 627–630.

[64] ERDOS, P., STEEL, M., SZEKELY, L., AND WARNOW, T. A few logs suffice to build (almost) all trees: Part II. *Theoretical Computer Science 221*, 1-2 (1999), 77–118.

[65] ERDOS, P. L., STEEL, M. A., SZÉKELY, L. A., AND WARNOW, T. J. A few logs suffice to build (almost) all trees (I). *Random Structures and Algorithms* (1999).

[66] FEDUCCIA, A. 'Big bang' for tertiary birds?, 2003.

[67] FELSENSTEIN, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution 17*, 6 (1981), 368–376.

[68] FELSENSTEIN, J. *Inferring phylogenies*. Sinauer Associates, Sunderland (MA), 2004.

[69] FELSENSTEIN, J., AND KISHINO, H. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Systematic Biology 42*, 2 (1993), 193–200.

[70] FELSENSTEIN JOSEPH. Confidence Limits on Phylogenies: an Approach Using the Bootstrap. *Evolution* (1985).

[71] FENG, Q., LIANG, S., JIA, H., STADLMAYR, A., TANG, L., LAN, Z., ZHANG, D., XIA, H., XU, X., JIE, Z., SU, L., LI, X., LI, X., LI, J., XIAO, L., HUBER-SCHÖNAUER, U., NIEDERSEER, D., XU, X., AL-AAMA, J. Y., YANG, H., WANG, J., KRISTIANSEN, K., ARUMUGAM, M., TILG, H., DATZ, C., AND WANG, J. Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nature Communications 6*, 1 (2015), 6528.

[72] FLETCHER, W., AND YANG, Z. INDELible: A flexible simulator of biological sequence evolution. *Molecular Biology and Evolution 26*, 8 (2009), 1879–1888.

[73] FLORES, G. E., CAPORASO, J. G., HENLEY, J. B., RIDEOUT, J. R., DOMOGALA, D., CHASE, J., LEFF, J. W., VÁZQUEZ-BAEZA, Y., GONZALEZ, A., KNIGHT, R., DUNN, R. R., AND FIERER, N. Temporal variability is a personalized feature of the human microbiome. *Genome Biology 15*, 12 (2014), 531.

[74] GALTIER, N., AND DAUBIN, V. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences* (2008).

[75] GATESY, J., AND SPRINGER, M. S. Phylogenetic Analysis at Deep Timescales: Unreliable Gene Trees, Bypassed Hidden Support, and the Coalescence/Concatalescence Conundrum. *Molecular phylogenetics and evolution 80* (2014), 231–266.

[76] GEE, H. Evolution: ending incongruence. *Nature 425*, 6960 (2003), 782.

[77] GEVERS, D., KUGATHASAN, S., DENSON, L. A., VÁZQUEZ-BAEZA, Y., VAN TREUREN, W., REN, B., SCHWAGER, E., KNIGHTS, D., SONG, S. J., YASSOUR, M., MORGAN, X. C., KOSTIC, A. D., LUO, C., GONZÁLEZ, A., MCDONALD, D., HABERMAN, Y., WALTERS, T., BAKER, S., ROSH, J., STEPHENS, M., HEYMAN, M., MARKOWITZ, J., BALDASSANO, R., GRIFFITHS, A., SYLVESTER, F., MACK, D., KIM, S., CRANDALL, W., HYAMS, J., HUTTENHOWER, C., KNIGHT, R., AND XAVIER, R. J. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host and Microbe* (2014).

[78] GIARLA, T. C., AND ESSELSTYN, J. A. The Challenges of Resolving a Rapid, Recent Radiation: Empirical and Simulated Phylogenomics of Philippine Shrews. *Systematic biology* (2015), syv029.

[79] GILL, S. R., POP, M., DEBOY, R. T., ECKBURG, P. B., TURNBAUGH, P. J., SAMUEL, B. S., GORDON, J. I., RELMAN, D. A., FRASER-LIGGETT, C. M., AND NELSON, K. E. Metagenomic Analysis of the Human Distal Gut Microbiome. *Science 312*, 5778 (2006), 1355–1359.

[80] GONZALEZ, A., NAVAS-MOLINA, J. A., KOSCIOLEK, T., MCDONALD, D., VÁZQUEZ-BAEZA, Y., ACKERMANN, G., DEREUS, J., JANSSEN, S., SWAFFORD, A. D., ORCHANIAN, S. B., SANDERS, J. G., SHORENSTEIN, J., HOLSTE, H., PETRUS, S., ROBBINS-PIANKA, A., BRISLAWN, C. J., WANG, M., RIDEOUT, J. R., BOLYEN, E., DILLON, M., CAPORASO, J. G., DORRESTEIN, P. C., AND KNIGHT, R. Qiita: rapid, web-enabled microbiome meta-analysis. *Nature Methods 15*, 10 (2018), 796–798.

[81] GOODMAN, S. A Dirty Dozen: Twelve P-Value Misconceptions. *Seminars in Hematology 45*, 3 (2008), 135–140.

[82] GOREMYKIN, V. V., NIKIFOROVA, S. V., BIGGS, P. J., ZHONG, B., DELANGE, P., MARTIN, W., WOETZEL, S., ATHERTON, R. A., MCLENACHAN, P. A., AND

LOCKHART, P. J. The Evolutionary Root of Flowering Plants. *Systematic Biology 62*, 1 (2013), 50–61.

[83] GRIMALDI, D., AND ENGEL, M. S. *Evolution of the Insects*. Cambridge University Press, 2005.

[84] GRIMALDI, D. A. 400 million years on six legs: on the origin and early evolution of Hexapoda. *Arthropod structure & development 39*, 2 (2010), 191–203.

[85] GROVER, C. E., GALLAGHER, J. P., JARECZEK, J. J., PAGE, J. T., UDALL, J. A., GORE, M. A., AND WENDEL, J. F. Re-evaluating the phylogeny of allopolyploid Gossypium L. *Molecular Phylogenetics and Evolution 92* (2015), 45–52.

[86] GUINDON, S., DELSUC, F., DUFAYARD, J.-F., AND GASCUEL, O. Estimating maximum likelihood phylogenies with PhyML. *Methods in Molecular Biology 537* (2009), 113–137.

[87] HAHN, M. W., AND NAKHLEH, L. Irrational exuberance for resolved species trees. *Evolution 70*, 1 (2016), 7–17.

[88] HASENFUSS, I. A possible evolutionary pathway to insect flight starting from lepismatid organization. *Journal of Zoological Systematics and Evolutionary Research 40*, 2 (6 2002), 65–81.

[89] HE, H., BAI, Y., GARCIA, E. A., AND LI, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (Hong Kong, 2008), pp. 1322–1328.

[90] HELED, J., AND DRUMMOND, A. J. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution 27*, 3 (2010), 570–580.

[91] HENNIG, W., HENNIG, W., HENNIG, W., ZOOLOGIST, G., AND HENNIG, W. Die stammesgeschichte der Insekten. *Frankfurt am Main* (1969).

[92] HERATY, J., RONQUIST, F., CARPENTER, J. M., HAWKS, D., SCHULMEISTER, S., DOWLING, A. P., MURRAY, D., MUNRO, J., WHEELER, W. C., SCHIFF, N., AND SHARKEY, M. Evolution of the hymenopteran megaradiation. *Molecular Phylogenetics and Evolution 60*, 1 (2011), 73–88.

[93] HILLIS, D. M., AND BULL, J. J. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Systematic Biology 42*, 2 (1993), 182–192.

[94] HILLIS, D. M., MORITZ, C., AND MABLE, B. K. *Molecular Systematics*, vol. 2nd. Sinauer Associates, 1996.

[95] Hoelzer, G. A., and Meinick, D. J. Patterns of speciation and limits to phylogenetic resolution. *Trends in Ecology & Evolution 9*, 3 (1994), 104–107.

[96] Hoschek, W. The Colt distribution: Open source libraries for high performance scientific and technical computing in JAVA, 2002.

[97] Hosner, P. A., Faircloth, B. C., Glenn, T. C., Braun, E. L., and Kimball, R. T. Avoiding Missing Data Biases in Phylogenomic Inference: An Empirical Study in the Landfowl (Aves: Galliformes). *Molecular Biology and Evolution 33*, 4 (2016), 1110–1125.

[98] Hovmöller, R., Knowles, L. L., and Kubatko, L. S. Effects of missing data on species tree estimation under the coalescent. *Molecular Phylogenetics and Evolution 69*, 3 (2013), 1057–1062.

[99] Huang, C.-H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L., Zhang, Q., Koch, M. A., Al-Shehbaz, I., and Edger, P. P. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular biology and evolution 33*, 2 (2016), 394–412.

[100] Huang, H., and Knowles, L. L. Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of rad sequences. *Systematic Biology 65*, 3 (2016), 357–365.

[101] Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics 14*, 1 (1998), 68–73.

[102] Ishiwata, K., Sasaki, G., Ogawa, J., Miyata, T., and Su, Z.-H. Phylogenetic relationships among insect orders based on three nuclear protein-coding gene sequences. *Molecular Phylogenetics and Evolution 58*, 2 (2011), 169–180.

[103] Jackman, T. R., Larson, A., de Queiroz, K., Losos, J. B., and Cannatella, D. Phylogenetic Relationships and Tempo of Early Diversification in Anolis Lizards. *Systematic Biology 48*, 2 (1999), 254–285.

[104] Janssen, S., McDonald, D., Gonzalez, A., Navas-Molina, J. A., Jiang, L., Xu, Z. Z., Winker, K., Kado, D. M., Orwoll, E., Manary, M., Mirarab, S., and Knight, R. Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. *mSystems 3*, 3 (2018).

[105] Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M. S., Zavidovych, V., Subramanian, S., Gabaldon, T., Capella-Gutierrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W. C., Ray, D., Green,

R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V., Lovell, P. V., Wirthlin, M., Schneider, M. P. C., Prosdocimi, F., Samaniego, J. A., Velazquez, A. M. V., Alfaro-Nunez, A., Campos, P. F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P., Driskell, A. C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F. E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K. K., Jonsson, K. A., Johnson, W., Koepfli, K.-P., O'Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R., Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alstrom, P., Edwards, S. V., Stamatakis, A., Mindell, D. P., Cracraft, J., Braun, E. L., Warnow, T., Jun, W., Gilbert, M. T. P. T. P., and Zhang, G. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science 346*, 6215 (2014), 1320–1331.

[106] Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. Phylogenomics: the beginning of incongruence? *Trends in Genetics 22*, 4 (2006), 225–231.

[107] Jiang, T., Kearney, P., and Li, M. A Polynomial Time Approximation Scheme for Inferring Evolutionary Trees from Quartet Topologies and Its Application, 2001.

[108] Jones, D. T., Taylor, W. R., and Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics 8*, 3 (1992), 275–282.

[109] Joseph, L., and Buchanan, K. L. A quantum leap in avian biology. *Emu - Austral Ornithology 115*, 1 (2015), 1–5.

[110] Junier, T., and Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics 26*, 13 (2010), 1669–1670.

[111] Kimball, R. T., Wang, N., Heimer-McGinn, V., Ferguson, C., and Braun, E. L. Identifying localized biases in large datasets: a case study using the avian tree of life. *Molecular Phylogenetics and Evolution 69*, 3 (2013), 1021–1032.

[112] Kjer, K. M., Carle, F. L., Litman, J., and Ware, J. A molecular phylogeny of Hexapoda. *Arthropod Syst Phylogeny 64*, 1 (2006), 35–44.

[113] Klass, K.-D. A Critical Review of Current Data and Hypotheses on Hexapod Phylogeny. In *Proc. Arthropod. Embryol. Soc. Jpn* (2009), vol. 43, pp. 3–22.

[114] Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C., and Knight, R. Human-Associated Microbial Signatures: Examining Their Predictive Value. *Cell Host & Microbe 10*, 4 (2011), 292–296.

[115] KOEHLER, K. J., AND LARNTZ, K. An Empirical Investigation of Goodness-of-Fit Statistics for Sparse Multinomials. *Journal of the American Statistical Association 75*, 370 (6 1980), 336–344.

[116] KRICHEVSKY, R. E., AND TROFIMOV, V. K. The Performance of Universal Encoding. *IEEE Transactions on Information Theory* (1981).

[117] KRISTENSEN, N. P. The phylogeny of hexapod "orders". A critical review of recent accounts. *Journal of Zoological Systematics and Evolutionary Research 13*, 1 (4 1975), 1–44.

[118] KRISTENSEN, N. P. Phylogeny of extant hexapods. In *The Insects of Australia*. Melbourne University Publishing, 1991, pp. 126–140.

[119] KRISTENSEN, N. P. Phylogeny of endopterygote insects, the most successful lineage of living organisms. *EJE 96*, 3 (1999), 237–253.

[120] KRISTENSEN, N. P., SCOBLE, M. J., AND KARSHOLT, O. L. E. Lepidoptera phylogeny and systematics: the state of inventorying moth and butterfly diversity. *Zootaxa 1668*, 699 (2007), e747.

[121] KUBAT, M., AND MATWIN, S. Addressing the curse of imbalanced training sets: One sided selection. In *Proceedings of the 14th International Conference on Machine Learning* (Nashville, Tennesse, 1997), p. 179–186.

[122] KUBATKO, L. S., AND DEGNAN, J. H. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology 56* (2007), 17–24.

[123] LANGILLE, M., ZANEVELD, J., CAPORASO, J. G., MCDONALD, D., KNIGHTS, D., REYES, J., CLEMENTE, J., BURKEPILE, D., VEGA THURBER, R., KNIGHT, R., BEIKO, R., AND HUTTENHOWER, C. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature biotechnology 31*, 9 (2013), 814–821.

[124] LARGET, B., AND SIMON, D. L. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* (1999).

[125] LARGET, B. R., KOTHA, S. K., DEWEY, C. N., AND ANÉ, C. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics 26*, 22 (2010), 2910–2911.

[126] LAUMER, C. E., HEJNOL, A., AND GIRIBET, G. Nuclear genomic signals of the 'microturbellarian' roots of platyhelminth evolutionary innovation. *eLife 4* (2015).

[127] LEEK, J. T., SCHARPF, R. B., BRAVO, H. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GEMAN, D., BAGGERLY, K., AND IRIZARRY, R. A. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics 11*, 10 (2010), 733–739.

[128] LEFORT, V., DESPER, R., AND GASCUEL, O. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program: Table 1. *Molecular Biology and Evolution 32*, 10 (2015), 2798–2800.

[129] LEMAÎTRE, G., NOGUEIRA, F., AND ARIDAS, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research 18*, 17 (2017), 1–5.

[130] LEMMON, A. R., BROWN, J. M., STANGER-HALL, K., AND LEMMON, E. M. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Systematic biology 58*, 1 (2 2009), 130–145.

[131] LEMMON, E. M., AND LEMMON, A. R. High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics 44*, 1 (2013), 99–121.

[132] LETSCH, H., AND SIMON, S. Insect phylogenomics: New insights on the relationships of lower neopteran orders (Polyneoptera). *Systematic Entomology 38*, 4 (10 2013), 783–793.

[133] LEWIS, P. O., HOLDER, M. T., AND HOLSINGER, K. E. Polytomies and Bayesian phylogenetic inference. *Systematic Biology 54*, 2 (2005), 241–253.

[134] LIAW, A., AND WIENER, M. Classification and Regression by randomForest. *R news* (2002).

[135] LINDER, C. R., AND WARNOW, T. An overview of phylogeny reconstruction. *Handbook of Computational Biology* (2005).

[136] LIU, B., GIBBONS, T., GHODSI, M., AND POP, M. MetaPhyler: Taxonomic profiling for metagenomic sequences. In *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on* (2011), IEEE, p. 95–100.

[137] LIU, G. E., ALKAN, C., JIANG, L., ZHAO, S., AND EICHLER, E. E. Comparative analysis of Alu repeats in primate genomes. *Genome Research 19*, 5 (2009), 876–885.

[138] LIU, K., LINDER, C. R., AND WARNOW, T. RAxML and FastTree: Comparing Two Methods for Large-Scale Maximum Likelihood Phylogeny Estimation. *PLoS ONE 6*, 11 (2011), e27731.

[139] LIU, L. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics 24*, 21 (2008), 2542–2543.

[140] LIU, L., XI, Z., AND DAVIS, C. C. Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. *Molecular Biology and Evolution 32*, 3 (2015), 791–805.

[141] LIU, L., AND YU, L. Estimating species trees from unrooted gene trees. *Systematic Biology 60*, 5 (2011), 661–667.

[142] LIU, L., YU, L., AND EDWARDS, S. V. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology 10*, 1 (2010), 302.

[143] LIU, L., YU, L., PEARL, D. K., AND EDWARDS, S. V. Estimating species phylogenies using coalescence times among sequences. *Systematic Biology 58*, 5 (2009), 468–477.

[144] LOZUPONE, C., AND KNIGHT, R. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology 71*, 12 (12 2005), 8228 LP – 8235.

[145] LOZUPONE, C. A., HAMADY, M., KELLEY, S. T., AND KNIGHT, R. Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. *Applied and Environmental Microbiology 73*, 5 (3 2007), 1576 LP – 1585.

[146] MADDISON, W. Reconstructing character evolution on polytomous cladograms. *Cladistics 5*, 4 (1989), 365–377.

[147] MADDISON, W. Gene Trees in Species Trees. *Systematic Biology 46*, 3 (1997), 523–536.

[148] MAI, U., AND MIRARAB, S. TreeShrink: Efficient Detection of Outlier Tree Leaves. In *Comparative Genomics: 15th International Workshop, RECOMB CG 2017, Barcelona, Spain, October 4-6, 2017, Proceedings*, J. Meidanis and L. Nakhleh, Eds. Springer International Publishing, Cham, 2017, pp. 116–140.

[149] MALLO, D., DE OLIVEIRA MARTINS, L., POSADA, D., MARTINS, L. D. O., AND POSADA, D. SimPhy: Phylogenomic Simulation of Gene, Locus and Species Trees. *Systematic biology 65*, 2 (6 2016), syv082–.

[150] MATSEN, F. A. Phylogenetics and the human microbiome. *Systematic Biology 64*, 1 (2014), e26–e41.

[151] MAU, B., NEWTON, M. A., AND LARGET, B. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* (1999).

[152] MCCORMACK, J. E., HARVEY, M. G., FAIRCLOTH, B. C., CRAWFORD, N. G., GLENN, T. C., AND BRUMFIELD, R. T. A Phylogeny of Birds Based on Over 1,500 Loci Collected by Target Enrichment and High-Throughput Sequencing. *PLoS ONE 8*, 1 (2013), e54848.

[153] McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., Aksenov, A. A., Behsaz, B., Brennan, C., Chen, Y., DeRight Goldasich, L., Dorrestein, P. C., Dunn, R. R., Fahimipour, A. K., Gaffney, J., Gilbert, J. A., Gogul, G., Green, J. L., Hugenholtz, P., Humphrey, G., Huttenhower, C., Jackson, M. A., Janssen, S., Jeste, D. V., Jiang, L., Kelley, S. T., Knights, D., Kosciolek, T., Ladau, J., Leach, J., Marotz, C., Meleshko, D., Melnik, A. V., Metcalf, J. L., Mohimani, H., Montassier, E., Navas-Molina, J., Nguyen, T. T., Peddada, S., Pevzner, P., Pollard, K. S., Rahnavard, G., Robbins-Pianka, A., Sangwan, N., Shorenstein, J., Smarr, L., Song, S. J., Spector, T., Swafford, A. D., Thackray, V. G., Thompson, L. R., Tripathi, A., Vázquez-Baeza, Y., Vrbanac, A., Wischmeyer, P., Wolfe, E., Zhu, Q., and Knight, R. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems 3*, 3 (2018), 00031–18.

[154] McDonald, D., Knight, R., Wendel, D., Stombaugh, J., Clemente, J. C., Kuczynski, J., Rideout, J. R., Caporaso, J. G., Wilke, A., Meyer, F., Hufnagle, J., and Huse, S. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience 1*, 1 (7 2012), 7.

[155] McKenna, D. D., and Farrell, B. D. 9-genes reinforce the phylogeny of holometabola and yield alternate views on the phylogenetic placement of Strepsiptera. *PLoS ONE 5*, 7 (7 2010), e11887.

[156] McMurdie, P. J., and Holmes, S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Computational Biology 10*, 4 (4 2014), e1003531.

[157] Metzker, M. L. Sequencing technologies - the next generation. *Nature Reviews Genetics 11*, 1 (2010), 31–46.

[158] Meusemann, K., von Reumont, B. M., Simon, S., Roeding, F., Strauss, S., Kück, P., Ebersberger, I., Walzl, M., Pass, G., Breuers, S., Achter, V., von Haeseler, A., Burmester, T., Hadrys, H., Wägele, J. W., and Misof, B. A Phylogenomic Approach to Resolve the Arthropod Tree of Life. *Molecular Biology and Evolution 27*, 11 (11 2010), 2451–2464.

[159] Mirarab, S., Bayzid, M. S., and Warnow, T. Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting. *Systematic Biology 0*, 0 (2014), 1–15.

[160] Mirarab, S., Bayzid, M. S., and Warnow, T. Evaluating summary methods for multi-locus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology 65*, 3 (2016), 366–380.

[161] Mirarab, S., Bayzid, S. M., Boussau, B., and Warnow, T. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science 346*, 6215 (2014), 1250463.

[162] Mirarab, S., Nguyen, N., and Warnow, T. SEPP: SATé-Enabled Phylogenetic Placement. In *Pacific Symposium On Biocomputing* (Fairmont Orchid, Big Island of Hawaii, 2012), WORLD SCIENTIFIC, pp. 247–258.

[163] Mirarab, S., Nguyen, N., and Warnow, T. PASTA: ultra-large multiple sequence alignment. *Research in Computational Molecular Biology* (2014), 177–191.

[164] Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics 30*, 17 (2014), i541–i548.

[165] Mirarab, S., and Warnow, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics 31*, 12 (6 2015), i44–i52.

[166] Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T., Beutel, R. G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A. J., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T. R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L. S., Kawahara, A. Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D. D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J. L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von Reumont, B. M., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N. U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M. G., Wiegmann, B. M., Wilbrandt, J., Wipfler, B., Wong, T. K. F., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D. K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K. M., and Zhou, X. Phylogenomics resolves the timing and pattern of insect evolution. *Science 346*, 6210 (11 2014), 763 LP – 767.

[167] Misof, B., Niehuis, O., Bischoff, I., Rickert, A., Erpenbeck, D., and Staniczek, A. Towards an 18S phylogeny of hexapods: Accounting for group-specific character covariance in optimized mixed nucleotide/doublet models. *Zoology 110*, 5 (2007), 409–429.

[168] MORTON, J. T., SANDERS, J., QUINN, R. A., MCDONALD, D., GONZALEZ, A., VÁZQUEZ-BAEZA, Y., NAVAS-MOLINA, J. A., SONG, S. J., METCALF, J. L., HYDE, E. R., LLADSER, M., DORRESTEIN, P. C., AND KNIGHT, R. Balance Trees Reveal Microbial Niche Differentiation. *mSystems 2*, 1 (2017), 00162–16.

[169] MOSSEL, E., AND ROCH, S. Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 7*, 1 (2010), 166–171.

[170] NAGARAJAN, N., AND POP, M. Sequence assembly demystified. *Nature Reviews Genetics 14*, 3 (2013), 157–167.

[171] NAKHLEH, L. Evolutionary Phylogenetic Networks: Models and Issues. *Networks* (2011), 125–158.

[172] NATIONAL RESEARCH COUNCIL (US) COMMITTEE ON METAGENOMICS: CHALLENGES AND FUNCTIONAL APPLICATIONS. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet.* Natl Academy Pr, 2007.

[173] NGUYEN, N., MIRARAB, S., LIU, B., POP, M., AND WARNOW, T. TIPP: Taxonomic Identification and Phylogenetic Profiling. *Bioinformatics 30*, 24 (2014), 3548–3555.

[174] NGUYEN, N., MIRARAB, S., AND WARNOW, T. MRL and SuperFine+ MRL: new supertree methods. *Algorithms for Molecular Biology 7*, 1 (2012), 3.

[175] NGUYEN, N.-P. D., MIRARAB, S., KUMAR, K., AND WARNOW, T. Ultra-large alignments using phylogeny-aware profiles. *Genome Biology 16*, 1 (2015), 124.

[176] NIEHUIS, O., HARTIG, G., GRATH, S., POHL, H., LEHMANN, J., TAFER, H., DONATH, A., KRAUSS, V., EISENHARDT, C., HERTEL, J., PETERSEN, M., MAYER, C., MEUSEMANN, K., PETERS, R. S., STADLER, P. F., BEUTEL, R. G., BORNBERG-BAUER, E., MCKENNA, D. D., AND MISOF, B. Genomic and morphological evidence converge to resolve the enigma of strepsiptera. *Current Biology 22*, 14 (2012), 1309–1313.

[177] O'DWYER, J. P., KEMBEL, S. W., AND GREEN, J. L. Phylogenetic Diversity Theory Sheds Light on the Structure of Microbial Communities. *PLoS Computational Biology 8*, 12 (2012), e1002832.

[178] PAMILO, P., AND NEI, M. Relationships between gene trees and species trees. *Mol Biol Evol 5*, 5 (1988), 568–583.

[179] PARADIS, E., CLAUDE, J., AND STRIMMER, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics 20*, 2 (2004), 289–290.

[180] PATEL, S., KIMBALL, R. T., AND BRAUN, E. L. Error in phylogenetic estimation for bushes in the Tree of Life. *Journal of Phylogenetics and {...} 1*, 2 (2013), 110.

[181] PAULSON, J. N., STINE, O. C., BRAVO, H. C., AND POP, M. Differential abundance analysis for microbial marker-gene surveys. *Nature methods 10*, 12 (2013), 1200–1202.

[182] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[183] PHILIPPE, H., BRINKMANN, H., LAVROV, D. V., LITTLEWOOD, D. T. J., MANUEL, M., WÖRHEIDE, G., AND BAURAIN, D. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS biology 9*, 3 (2011), e1000602.

[184] PHILIPPE, H., SNELL, E. A., BAPTESTE, E., LOPEZ, P., HOLLAND, P. W. H., AND CASANE, D. Phylogenomics of Eukaryotes: Impact of Missing Data on Large Alignments. *Molecular Biology and Evolution 21*, 9 (9 2004), 1740–1752.

[185] PHILLIPS, M. J., DELSUC, F., AND PENNY, D. Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution* (2004).

[186] PIAGGIO-TALICE, R., BURLEIGH, J. G., AND EULENSTEIN, O. Quartet Supertrees. In *Phylogenetic Supertrees SE - 9*, O. Bininda-Emonds, Ed., vol. 4 of *Computational Biology*. Springer Netherlands, 2004, pp. 173–191.

[187] POE, S., AND CHUBB, A. L. Birds in a bush: five genes indicate explosive evolution of avian orders. *Evolution; international journal of organic evolution 58*, 2 (2004), 404–415.

[188] PRICE, M. N., DEHAL, P. S., AND ARKIN, A. P. FastTree-2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE 5*, 3 (2010), e9490.

[189] PRUM, R. O., BERV, J. S., DORNBURG, A., FIELD, D. J., TOWNSEND, J. P., LEMMON, E. M., AND LEMMON, A. R. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature advance on* (10 2015).

[190] RAGAN, M. A. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution 1*, 1 (1992), 53–58.

[191] RANNALA, B. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics 1656*, August (2003), 1645–1656.

[192] RANNALA, B., AND YANG, Z. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution* (1996).

[193] READ, T. R. C., AND CRESSIE, N. A. C. *Goodness-of-fit statistics for discrete multivariate data*. Springer Science & Business Media, 2012.

[194] Regier, J. C., Shultz, J. W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J. W., and Cunningham, C. W. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature 463*, 7284 (2 2010), 1079–1083.

[195] Richards, S., and Murali, S. C. Best Practices in Insect Genome Sequencing: What Works and What Doesn't Sanger Beginnings: The First Insect Genome. *Current Opinion in Insect Science 7* (2015), 1–7.

[196] Robinson, D. F., and Foulds, L. R. Comparison of phylogenetic trees. *Mathematical Biosciences 53*, 1-2 (1981), 131–147.

[197] Roch, S., and Snir, S. Recovering the Treelike Trend of Evolution Despite Extensive Lateral Genetic Transfer: A Probabilistic Analysis. *Journal of Computational Biology 20*, 2 (2013), 93–112.

[198] Roch, S., and Steel, M. M. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical population biology 100* (2014), 56–62.

[199] Roch, S., and Warnow, T. On the Robustness to Gene Tree Estimation Error (or lack thereof) of Coalescent-Based Species Tree Methods. *Systematic Biology 64*, 4 (2015), 663–676.

[200] Rokas, A., Williams, B. L., King, N., and Carroll, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature 425*, 6960 (2003), 798–804.

[201] Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N., and Douzery, E. J. P. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Molecular biology and evolution 30*, 9 (2013), 2134–2144.

[202] Rosenberg, N. A. Discordance of species trees with their most likely gene trees: a unifying principle. *Molecular Biology and Evolution 30*, 12 (2013), 2709–2713.

[203] Rouse, G. W., Wilson, N. G., Carvajal, J. I., and Vrijenhoek, R. C. New deep-sea species of Xenoturbella and the position of Xenacoelomorpha. *Nature 530*, 7588 (2016), 94–97.

[204] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV) 115*, 3 (2015), 211–252.

[205] Russell, D. J. *Multiple sequence alignment methods*. Springer, 2014.

[206] SAITOU, N., AND NEI, M. The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution 4*, 4 (1987), 406–425.

[207] SALICHOS, L., AND ROKAS, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature 497*, 7449 (2013), 327–331.

[208] SAULNIER, D. M., RIEHLE, K., MISTRETTA, T., DIAZ, M., MANDAL, D., RAZA, S., WEIDLER, E. M., QIN, X., COARFA, C., MILOSAVLJEVIC, A., PETROSINO, J. F., HIGHLANDER, S., GIBBS, R., LYNCH, S. V., SHULMAN, R. J., AND VERSALOVIC, J. Gastrointestinal Microbiome Signatures of Pediatric Patients With Irritable Bowel Syndrome. *Gastroenterology 141*, 5 (2011), 1782–1791.

[209] SAVARD, J., TAUTZ, D., RICHARDS, S., WEINSTOCK, G. M., GIBBS, R. A., WERREN, J. H., TETTELIN, H., AND LERCHER, M. J. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Research 16*, 11 (11 2006), 1334–1338.

[210] SAYYARI, E., AND MIRARAB, S. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Molecular Biology and Evolution 33*, 7 (2016), 1654–1668.

[211] SAYYARI, E., AND MIRARAB, S. Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes 9*, 3 (8 2018), 132.

[212] SAYYARI, E., WHITFIELD, J. B., AND MIRARAB, S. Fragmentary Gene Sequences Negatively Impact Gene Tree and Species Tree Reconstruction. *Molecular Biology and Evolution 34*, 12 (2017), 3279–3291.

[213] SCHLOSS, P. D., AND HANDELSMAN, J. Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Applied and Environmental Microbiology 71*, 3 (2005), 1501–1506.

[214] SEO, T. K. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molecular Biology and Evolution 25*, 5 (2008), 960–971.

[215] SEO, T.-K., KISHINO, H., AND THORNE, J. L. Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. *Proceedings of the National Academy of Sciences 102*, 12 (2005), 4436–4441.

[216] SHEKHAR, S., ROCH, S., AND MIRARAB, S. Species Tree Estimation Using ASTRAL: How Many Genes Are Enough? *IEEE/ACM Transactions on Computational Biology and Bioinformatics 15*, 5 (2018), 1738–1747.

[217] SHEN, X. X., HITTINGER, C. T., AND ROKAS, A. Studies Can Be Driven By a Handful of Genes. *Nature 1*, April (2017), 1–10.

[218] SIMMONS, M. P. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. *Cladistics 28*, 2 (2012), 208–222.

[219] SIMMONS, M. P. A confounding effect of missing data on character conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses. *Molecular Phylogenetics and Evolution 80*, 1 (2014), 267–280.

[220] SIMMONS, M. P., AND GATESY, J. Coalescence vs. concatenation: sophisticated analyses vs. first principles applied to rooting the angiosperms. *Molecular phylogenetics and evolution 91* (2015), 98–122.

[221] SIMMONS, M. P., SLOAN, D. B., SPRINGER, M. S., AND GATESY, J. Gene-wise resampling outperforms site-wise resampling in phylogenetic coalescence analyses. *Molecular Phylogenetics and Evolution* (2019).

[222] SLOWINSKI, J. B. Molecular Polytomies. *Molecular Phylogenetics and Evolution 19*, 1 (2001), 114–120.

[223] SNIR, S., AND RAO, S. Quartets MaxCut: A divide and conquer quartets algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 7*, 4 (2010), 704–718.

[224] SOKAL, R. R., AND MICHENER, C. *A statistical method for evaluating systematic relationships*. University of Kansas, 1958.

[225] SOLÍS-LEMUS, C., YANG, M., AND ANÉ, C. Inconsistency of Species Tree Methods under Gene Flow. *Systematic Biology 65*, 5 (2016), 843–851.

[226] SOLTIS, D. E., SMITH, S. A., CELLINESE, N., WURDACK, K. J., TANK, D. C., BROCKINGTON, S. F., REFULIO-RODRIGUEZ, N. F., WALKER, J. B., MOORE, M. J., CARLSWARD, B. S., BELL, C. D., LATVIS, M., CRAWLEY, S., BLACK, C., DIOUF, D., XI, Z., RUSHWORTH, C. A., GITZENDANNER, M. A., SYTSMA, K. J., QIU, Y., HILU, K. W., DAVIS, C. C., SANDERSON, M. J., BEAMAN, R. S., OLMSTEAD, R. G., JUDD, W. S., DONOGHUE, M. J., AND SOLTIS, P. S. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany 98* (2011), 704–730.

[227] SONG, S., LIU, L., EDWARDS, S. V., AND WU, S. Correction for Song et al., Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences 112*, 44 (2012), E6079–E6079.

[228] SPRINGER, M. S., AND GATESY, J. Land plant origins and coalescence confusion. *Trends in plant science 19*, 5 (2014), 267–269.

[229] SPRINGER, M. S., AND GATESY, J. The gene tree delusion. *Molecular Phylogenetics and Evolution 94*, Part A (2016), 1–33.

[230] SPRINGER, M. S., AND GATESY, J. On the importance of homology in the age of phylogenomics. *Systematics and Biodiversity* (2017), 1–19.

[231] STADLER, T., AND STEEL, M. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. *Journal of Theoretical Biology 297* (2012), 33–40.

[232] STAMATAKIS, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics 30*, 9 (2014), 1312–1313.

[233] STATNIKOV, A., HENAFF, M., NARENDRA, V., KONGANTI, K., LI, Z., YANG, L., PEI, Z., BLASER, M. J., ALIFERIS, C. F., AND ALEKSEYENKO, A. V. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome 1*, 1 (2013), 11.

[234] STEEL, M. Recovering a tree from the leaf colourations it generates under a Markov model. *Applied Mathematics Letters 7*, 2 (1994), 19–23.

[235] STENZ, N. W. M., LARGET, B. R., BAUM, D. A., AND ANÉ, C. Exploring Tree-Like and Non-Tree-Like Patterns Using Genome Sequences: An Example Using the Inbreeding Plant Species ¡i¿Arabidopsis thaliana¡/i¿ (L.) Heynh. *Systematic Biology 64*, 5 (2015), 809–823.

[236] STREICHER, J. W., SCHULTE, J. A., AND WIENS, J. J. How Should Genes and Taxa be Sampled for Phylogenomic Analyses with Missing Data? An Empirical Study in Iguanian Lizards. *Systematic Biology 65*, 1 (2016), 128–145.

[237] STRIMMER, K., AND VON HAESELER, A. Quartet puzzling - a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular biology and evolution 13* (1996), 964–969.

[238] STUDIER, J. A., AND KEPPLER, K. J. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular biology and evolution 5*, 6 (1988), 729–731.

[239] SUH, A. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. *Zoologica Scripta 45* (2016), 50–62.

[240] SUH, A., PAUS, M., KIEFMANN, M., CHURAKOV, G., FRANKE, F. A., BROSIUS, J., KRIEGS, J. O., AND SCHMITZ, J. Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. *Nature communications 2* (2011), 443.

[241] SUH, A., SMEDS, L., AND ELLEGREN, H. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biol 13*, 8 (2015), e1002224.

[242] SUKUMARAN, J., AND HOLDER, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics 26*, 12 (2010), 1569–1571.

[243] SUSKO, E. Bootstrap support is not first-order correct. *Systematic Biology 58*, 2 (2009), 211–223.

[244] SWENSON, M. S., SURI, R., LINDER, C. R., AND WARNOW, T. SuperFine: fast and accurate supertree estimation. *Systematic biology 61*, 2 (2012), 214–227.

[245] SWOFFORD, D. L. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4., 2003.

[246] SWOFFORD, D. L., OLSEN, G. J., WADDELL, P. J., AND HILLIS, D. M. Phylogenetic Inference. In *Molecular systematics*. Sinauer, Sunderland, Massachusetts, 1996, p. 655.

[247] SZE, M. A., AND SCHLOSS, P. D. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. *mBio 7*, 4 (2016), 01018–16.

[248] SZÖLLÕSI, G. J., TANNIER, E., DAUBIN, V., AND BOUSSAU, B. The inference of gene trees with species trees. *Systematic Biology 64*, 1 (2014), e42–e62.

[249] TAVARÉ, S. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences 17* (1986), 57–86.

[250] TERRY, M. D., AND WHITING, M. F. Mantophasmatodea and phylogeny of the lower neopterous insects. *Cladistics 21*, 3 (6 2005), 240–257.

[251] TOWNSEND, J. P., SU, Z., AND TEKLE, Y. I. Phylogenetic Signal and Noise: Predicting the Power of a Data Set to Resolve Phylogeny. *Systematic Biology 61*, 5 (2012), 835.

[252] TRAUTWEIN, M. D., WIEGMANN, B. M., BEUTEL, R., KJER, K. M., AND YEATES, D. K. Advances in Insect Phylogeny at the Dawn of the Postgenomic Era. *Annual Review of Entomology 57*, 1 (12 2012), 449–468.

[253] TRIPATHI, R. C., GUPTA, R. C., AND GURLAND, J. Estimation of parameters in the beta binomial model. *Annals of the Institute of Statistical Mathematics 46*, 2 (1994), 317–331.

[254] TURNBAUGH, P. J., LEY, R. E., HAMADY, M., FRASER-LIGGETT, C. M., KNIGHT, R., AND GORDON, J. I. The Human Microbiome Project. *Nature 449*, 7164 (2007), 804–810.

[255] VACHASPATI, P., AND WARNOW, T. ASTRID: Accurate Species TRees from Internode Distances. *BMC Genomics 16 Suppl 1*, Suppl 10 (2015), S3.

[256] VENTER, J. C., REMINGTON, K., HEIDELBERG, J. F., HALPERN, A. L., RUSCH, D., EISEN, J. A., WU, D., PAULSEN, I., NELSON, K. E., NELSON, W., FOUTS, D. E., LEVY, S., KNAP, A. H., LOMAS, M. W., NEALSON, K., WHITE, O., PETERSON, J., HOFFMAN, J., PARSONS, R., BADEN-TILLSON, H., PFANNKOCH, C., ROGERS,

Y.-H., and Smith, H. O. Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, N.Y.) 304*, 5667 (2004), 66–74.

[257] von Mering, C., Hugenholtz, P., Raes, J., Tringe, S. G., Doerks, T., Jensen, L. J., Ward, N., and Bork, P. Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments. *Science 315*, 5815 (2007), 1126–1130.

[258] Waldor, M. K., Tyson, G., Borenstein, E., Ochman, H., Moeller, A., Finlay, B. B., Kong, H. H., Gordon, J. I., Nelson, K. E., Dabbagh, K., and Smith, H. Where Next for Microbiome Research? *PLOS Biology 13*, 1 (2015), e1002050.

[259] Walsh, H. E., Kidd, M. G., Moum, T., and Friesen, V. L. Polytomies and the power of phylogenetic inference. *Evolution 53*, 3 (1999), 932–937.

[260] Warnow, T. Tree Compatibility and Inferring Evolutionary History. *Journal of Algorithms 16* (1994), 388–407.

[261] Weiss, S., Amir, A., Hyde, E. R., Metcalf, J. L., Song, S. J., and Knight, R. Tracking down the sources of experimental contamination in microbiome studies. *Genome Biology 15*, 12 (2014), 564.

[262] Wheat, C. W., and Wahlberg, N. Phylogenomic Insights into the Cambrian Explosion, the Colonization of Land and the Evolution of Flight in Arthropoda. *Systematic Biology 62*, 1 (2013), 93.

[263] Wheeler, T. J. Large-scale neighbor-joining with NINJA. In *Algorithms in Bioinformatics*. Springer, 2009, pp. 375–389.

[264] Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E. J., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., Ruhfel, B. R., Wafula, E., Der, J. P., Graham, S. W., Mathews, S., Melkonian, M., Soltis, D. E., Soltis, P. S., Miles, N. W., Rothfels, C. J., Pokorny, L., Shaw, A. J., DeGironimo, L., Stevenson, D. W., Surek, B., Villarreal, J. C., Roure, B., Philippe, H., DePamphilis, C. W., Chen, T., Deyholos, M. K., Baucom, R. S., Kutchan, T. M., Augustin, M. M., Wang, J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G. K.-S., and Leebens-Mack, J. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences 111*, 45 (2014), E4859–4868.

[265] Wickham, H. *ggplot2: elegant graphics for data analysis*. No. July in Use R! Springer International Publishing, Cham, 2016.

[266] WIEGMANN, B. M., TRAUTWEIN, M. D., KIM, J.-W., CASSEL, B. K., BERTONE, M. A., WINTERTON, S. L., AND YEATES, D. K. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. *BMC Biology 7* (6 2009), 34.

[267] WIENS, J. J. Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics 39*, 1 SPEC. ISS. (2 2006), 34–42.

[268] WIENS, J. J., AND MORRILL, M. C. Missing Data in Phylogenetic Analysis: Reconciling Results from Simulations and Empirical Data. *Systematic biology 60*, 5 (3 2011), 719–731.

[269] WIPFLER, B., MACHIDA, R., M??LLER, B., AND BEUTEL, R. G. On the head morphology of Grylloblattodea (Insecta) and the systematic position of the order, with a new nomenclature for the head muscles of Dicondylia. *Systematic Entomology 36*, 2 (4 2011), 241–266.

[270] WONG, G. K.-S. The Thousand Transcriptome (1KP) Project. {\textbackslash}url{http://www.onekp.com/project.html}, 2013.

[271] WOOD, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*, 1 (2011), 3–36.

[272] XI, Z., LIU, L., AND DAVIS, C. C. The impact of missing data on species tree estimation. *Molecular Biology and Evolution 33*, 3 (2016), 838–860.

[273] XI, Z., LIU, L., REST, J. S., AND DAVIS, C. C. Coalescent versus concatenation methods and the placement of Amborella as sister to water lilies. *Systematic Biology 63*, 6 (2014), 919–932.

[274] XU, B., AND YANG, Z. Challenges in species tree estimation under the multispecies coalescent model. *Genetics 204*, 4 (2016), 1353–1368.

[275] YOSHIZAWA, K. Monophyletic Polyneoptera recovered by wing base structure. *Systematic Entomology 36*, 3 (7 2011), 377–394.

[276] ZAR, J. H. *Biostatistical Analysis*. Pearson Education India, 2007.

[277] ZHANG, C., RABIEE, M., SAYYARI, E., AND MIRARAB, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics 19*, S6 (2018), 153.

[278] ZHANG, C., SAYYARI, E., AND MIRARAB, S. ASTRAL-III: Increased Scalability and Impacts of Contracting Low Support Branches. In *Lecture Notes in Computer Science*, J. Meidanis and L. Nakhleh, Eds., vol. 10562 LNBI. Springer International Publishing, Cham, 2017, pp. 53–75.

[279] ZHANG, N., ZENG, L., SHAN, H., AND MA, H. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. *New Phytologist 195* (2012), 923–937.

[280] ZIMMERMANN, T., MIRARAB, S., AND WARNOW, T. BBCA: Improving the scalability of *BEAST using random binning. *BMC genomics 15*, Suppl 6 (2014), S11.

[281] ZWICKL, D. J., STEIN, J. C., WING, R. A., WARE, D., AND SANDERSON, M. J. Disentangling methodological and biological sources of gene tree discordance on Oryza (Poaceae) chromosome 3. *Systematic Biology 63*, 5 (2014), 645–659.