Summer 7-5-2022

Heterogeneity of Gene Trees

Jonathan Nenye Odumegwu UNM
University of New Mexico - Main Campus

Jonathan Nenye Odumegwu

Candidate

Mathematics and Statistics

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Prof. James H. Degnan, Chair

Prof. Yan Lu, Member

Prof. Helen Wearing, Member

Prof. Joseph Cook, Member

Heterogeneity of Gene Trees

by

Jonathan Nenye Odumegwu

B.Sc., Mathematics, Nnamdi Azikiwe University, 2004

M.A., Mathematics, Central Michigan University, 2017

DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Statistics

The University of New Mexico

Albuquerque, New Mexico

July, 2022

Dedication

To my late mother.

Acknowledgments

I wish to thank my advisor, Professor James H. Degnan, for his invaluable help and supervision throughout the writing of this dissertation. His patience and step-by-step guidance made this dream come through. Without his competence and continuous mentoring, I wouldn't have been able to put this work together in phylogenetic studies. I would like to thank my committee members, Professors Yan Lu, Helen Wearing, and Joseph Cook, for their willingness to serve on my committee. Also, I am thankful to many Mathematics and Statistics faculty and staff, particularly Professor Maria Cristina, Dr. Janet Vassilev, and Ana ParaLombard, for their assistance and encouragement throughout my studentship in the department. Additionally, I am grateful to Dr. Anastesiia Kim, who introduced me to supercomputing/cluster machines and how to write bash scripts.

I would also like to thank my wife, Ekene and my children, Kamsi, Chinemelum, and Chidindu, for their support and patience during this study. Also, I express my gratitude and love to my brothers and sisters, particularly my elder sister, Virginia, for supporting me throughout my education. Without sister Virgy's devotion to my well-being and education, I wouldn't get to this point, especially after my mom's death. Further, I thank my friends for their love and goodwill.

Heterogeneity of Gene Trees

by

Jonathan Nenye Odumegwu

B.Sc., Mathematics, Nnamdi Azikiwe University, 2004

M.A., Mathematics, Central Michigan University, 2017

Ph.D., Statistics, University of New Mexico, 2022

Abstract

Multilocus phylogenetic studies often show a high degree of gene tree heterogeneity —gene trees that have different topologies from each other as well as from the species tree topology. In some cases, this can lead to studies with hundreds of loci having distinct gene tree topologies. The degree of heterogeneity is expected to increase when there is a high degree of incomplete lineage sorting due to short branches (as measured in coalescent units) in the species tree. Other potential sources of heterogeneity include other biological processes such as introgression, recombination within genes, ancestral population structure, gene duplication and loss, and horizontal gene transfer, as well as gene tree estimation error due to short DNA sequences or inadequate substitution models. Here we examine the relationships between speciation and extinction rates and gene tree heterogeneity with both gene tree estimation error and no gene tree estimation error. In particular, higher speciation rates lead to shorter branches in the species tree and, therefore, higher levels of incomplete lineage sorting. In many cases, it might not be surprising that every gene tree has a unique topology, even for data sets with 1000 gene trees. We also propose using the average pairwise Robinson-Foulds (RF) distance between gene trees as a measure of heterogeneity as opposed to using the average RF distance between gene trees and the true species tree. Further, methods of inferring birth-death parameters (speciation and extinction rates) have involved using species trees estimated from gene trees or concatenation of DNA sequences. We infer these parameters using gene trees instead of species trees in this work. The method uses Approximate Bayesian Computation (ABC), which is useful when the maximum likelihood method is intractable, as in the case of gene trees given a species tree with a large number of taxa.

Contents

Chapter 1

Introduction

1.1 Background

A phylogeny is a tree representing a group of species' evolutionary relationships and histories. The tree depicts how the lineages of the present day species descended from their ancestors. Life on earth has undergone and evolved several biological and environmental changes, and traces of these evolutionary histories have been active research areas in biological studies. For example, changes in traits (physically or otherwise) can occur among organisms of different species or organisms of the same species. For instance, American flamingos are pink (or reddish) due to carotenoids in their animal and plant plankton diet. In contrast, lesser flamingos (in India and Sub-Saharan Africa) are paler pink due to the small amount of carotenoids in their diets (Ali, 1990; McCulloch and Irvine, 2004; Anderson and Anderson, 2010). These changes in the organisms have several causal factors, particularly environmental and biological (genetic) factors. Sometimes, gradual changes lead to a divergence (speciation) of organisms, and identifying the modes of the divergence and how often it occurs is an active area of study in biology.

Further, species adaptations to their environments can lead to genetic changes in a population's history that can impact future generations. These histories can be studied and identified using phylogenetic trees. The main goal of phylogenetics is to reconstruct the evolutionary history of present day organisms and describe their historical relationship using a tree-like form. Of course, all organisms can be identified by their deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) for some viruses, and their evolutionary relationships can be studied by comparing their DNA sequences. DNA is made up of four types of nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G). Typically, a phylogenetic tree represents the evolutionary relationships of a set species given DNA sequences. The major cause of evolutionary change is mutations in the nucleotide sites.

A common approach to reconstruct the evolutionary histories is the concatenation of DNA sequences obtained from genes from single species at multiple genome sites. Concatenated sequences from multiple species can be aligned into a supermatrix (Chesters and Vogler, 2013). However, many studies (e.g., Kubatko and Degnan (2007); Leaché and Rannala (2011)) reported that such concatenation of sequences from multiple genes could bias the estimate of the true phylogeny (Leaché and Rannala, 2011).

1.1.1 Gene and Gene Trees

The term ***gene*** is used multiple ways in the literature, and here we use the molecular definition of gene — a contiguous sequence of DNA. A ***gene tree*** describes the evolutionary relationships between a sample of sequences for a non-recombining locus. Different gene trees can have separate evolutionary histories. Since every gene evolved differently, there is no reason to expect that gene trees based on different genes should have the same topologies. Two genes coalesce, going backward in time when their historical lineages merge into a single ancestral copy based on a random process. Gene trees of the genes sampled are contained within the branches of the species tree and represent the evolutionary histories of the genes. The method for computing gene tree probabilities involves keeping track of lineages from a gene tree coalescing on the branches of the species tree (Nei, 1987; Degnan and Rosenberg, 2006, 2009).

In this project, we focused on the gene tree topologies extracted from a bifurcating ultrametric species tree (the tips are equally distant from the root), see Figure 1.1. There are several types of tree topologies, rooted (ranked and unranked rooted) and unrooted gene trees. A gene tree is ***ranked*** if the order in which the lineages coalesce is important, and it is said to be ***unranked*** if the order is not important (see plots (a) and (b) on Figure 1.1). Using Newick format, a ***clade*** (a set of descendants from one ancestor) is represented by a pair of parentheses. We can write a ranked gene tree topology by modifying an unranked tree topology. For example, $((a,b)_2,(c,d)_3)$ and $((a,b)_3,(c,d)_2)$ are two different ranked tree topologies from the unranked tree topology, $((a,b),(c,d))$. In the tree, $((a,b)_2,(c,d)_3)$, the c and d lineages first coalesce (i.e., most recently), going back in time, before a and b coalesce, while (a,b) first coalesce before (c,d) in the second tree, $((a,b)_3,(c,d)_2)$. The largest subscript is the most recent coalescent event, and the rank of the root is 1, but is usually not stated.

The ***coalescent process*** is a stochastic model for the random joining of sampled lineages going backward in time (Kingman, 1982). Also, in the coalescent model, the time until a coalescent event has an exponential distribution, and its parameter depends on the remaining number of sampled lineages and the population size.

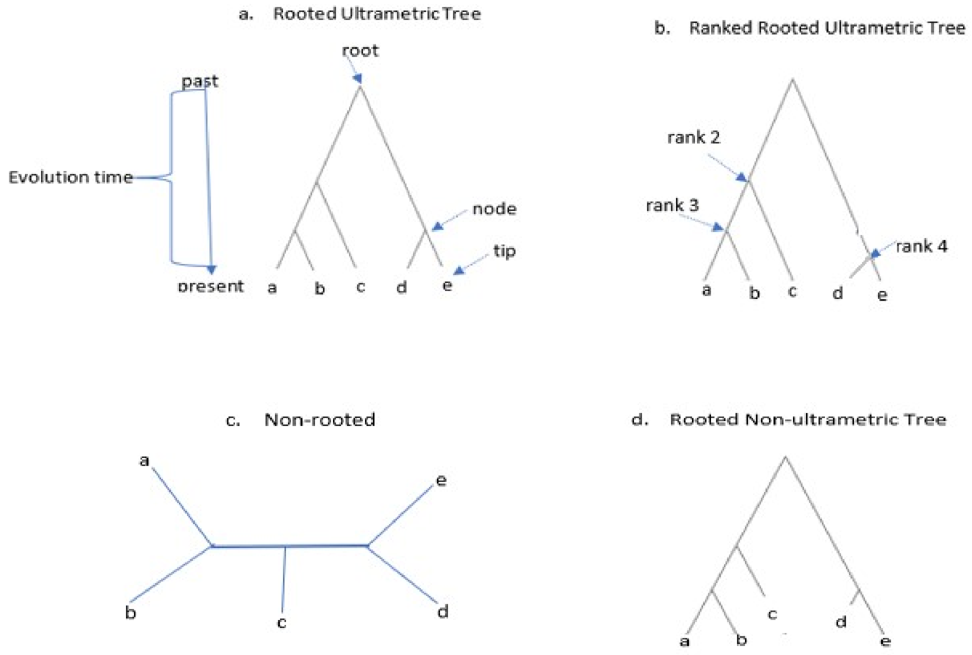
a. Rooted Ultrametric Tree
root
past
Evolution time
node
tip
present
a
b
c
d
e
b. Ranked Rooted Ultrametric Tree
rank 2
rank 3
rank 4
a
b
c
d
e
c. Non-rooted
a
e
b
c
d
d. Rooted Non-ultrametric Tree
a
b
c
d
e

Figure 1.1: Phylogenetic trees. (a), (b) and (d) are rooted and (c) is unrooted tree. Also, (a) -(b) are ultrametric, (a) has unranked topology $(((a, b), c),(d, e))$ and (b) has a ranked topology $(((a, b)_3, c)_2,(d, e)_4)$.

The coalescent model is drived from the classical population genetic model, the Fisher-Wright model (Fieher, 1922; Wright, 1931), which assumes that genetic differences between individuals do not influence their probability of reproducing; each gene copy is uniformly probable to have been passed from one generation to another. Also, populations with the same effective size (an ideal population where all nodes have an equal expectation of being the ancestors of any descendant(s)) have similar patterns of genetic variation and genetic drift (random fluctuations in allele frequency over time) as randomly mating populations regardless of the actual size of the population.

1.1.2 Species Tree Inference

A ***species tree*** describes the evolutionary relationships between a set of species. In phylogenetic and phylogenomic studies, the species tree is unknown and is treated as a parameter, while gene trees are treated as random variables whose distributions depend on the species tree. Based on the increase of the genomic data, species tree estimation from multiple loci sampled from different genomes is now on the rise but is challenged by the variability across the genome due to some biological processes, such as gene duplication and loss, horizontal gene transfer, and incomplete lineage sorting (Maddison, 1997; Rosenberg and Kumar, 2003; Degnan and Rosenberg, 2006, 2009; Molloy and Warnow, 2018). Many methods for estimating species trees have been developed to address gene tree heterogeneity due to incomplete lineage sorting and other biological processes (Degnan and Rosenberg, 2009). These methods combine estimated gene trees from multiple loci and are sensitive to gene tree variability and quality. These methods include minimizing deep coalescence MDC (Maddison, 1997; Maddison and Knowles, 2006), ASTRAL (Mirarab et al., 2014), ASTRID (Vachaspati and Warnow, 2015), and maximum likelihood-based methods like STEM (Kubatko et al., 2009). The non-parametric method BCA: Bayesian concordance approach (Ané et al., 2007) has been developed to handle different levels of gene tree uncertainty and to better infer species trees from gene trees. BCA (Ané et al.,

2007) infers the species tree by integrating over gene tree uncertainty and makes no assumption about the reason for discordance, and it uses a non-parametric clustering of genes with information sharing across compatible genes. Unfortunately, these methods have been impacted by gene tree estimation error, missing data, ILS, duplication and loss, and horizontal gene transfer (HGT) (Bansal et al., 2015; Molloy and Warnow, 2018; Bossert et al., 2021). The maximum likelihood procedure is used for reconstructing the phylogeny from concatenated DNA sequences. Still, it has been reported to be statistically inconsistent under the multispecies coalescent model due to its assumption that all genes share the same tree (Kubatko and Degnan, 2007; Roch and Steel, 2015). In addition, there are site-based methods such as SVDquartets (Chifman and Kubatko, 2014) and SNAPP (Bryant et al., 2012) that infer species tree directly from the site without the gene tree estimation.

1.1.3 Discordance of gene and species trees

Gene trees are used to reconstruct a species tree that describes evolutionary relationships among species. Gene trees that are contained within the branches of the species phylogeny represent the evolutionary histories of the sampled genes. Gene trees and species trees may or may not be the same; several processes can lead to discordance between gene trees and species trees and incongruence among the gene trees from a species tree. A common procedure for multilocus phylogeny estimation is concatenation, in which the DNA alignments are combined into a single supermatrix. However, different nucleotide sites evolving along the branches of a tree of species relationships can have different evolutionary histories, and consequently, estimates of species trees from genetic data can be influenced by the particular choice of nucleotides or genomic regions used in the analysis.

Another source of poor inference of species trees is over filtering of genes on the basis of missing data (Huang and Knowles, 2016; Molloy and Warnow, 2018) since removing genes based on the missing data from inference is likely to reduce the accuracy of the estimate rather than improve it. It has been suggested that methods that co-estimate gene trees and species seem to perform better than any other methods but are computationally expensive on large data sets (Molloy and Warnow, 2018). Development of new robust methods are needed to bridge the deficiencies of the existing methods. However, the contributions of the number of species, birth-death parameters (speciation and extinction rates), and sample size (number of loci) to the levels of gene trees heterogeneity and discordance of the species and gene trees have not been completely investigated.

In this project, we use two approaches: (1) an application of a generalized birthday problem to investigate the contributions of the sample size of gene trees (or loci) and the number of tips of the species tree, holding other biological factors constant, to the incongruence and discordance of gene and species trees; (2) simulation approach: we randomly generate the species tree using the birth-death process using TreeSim (Stadler, 2019), based on several combinations of parameters to investigate if the birth-death parameters are factors of the gene trees variability. We discuss these approaches in detail in the section below and in chapter 2.

1.1.4 Birth-Death (BD) Parameters

Birth-death (BD) processes are widely used to model the development of biological populations, and their parameters can be challenging to estimate because the likelihood can be computationally expensive or intractable as more data are added. A **birth-death model** is a continuous-time Markov process commonly used to study how the number of individuals in a population evolves through time. These individuals are usually species (lineages) in the case of macroevolution. A popular and common way of inference of the BD parameters is: first estimate gene trees from multiple genomic regions; second, estimate the species phylogeny from the gene trees; and third, estimate the BD parameters from the estimated species tree.

However, in the phylogenetic and phylogenomic studies, the true phylogeny is not available. Also, some biological processes such as hybridization and loss, horizontal gene transfer (HGT) and incomplete lineaging sorting (ILS), and gene trees estimation error (Molloy and Warnow, 2018) can seriously bias the estimates of the true phylogeny from the estimated gene trees and even directly from the genomic data. We propose estimating BD parameters directly from the gene trees using an Approximate Bayesian Computation framework based on these observations. We will be discussing this in detail in chapter 3.

Chapter 2

Heterogeneity of Gene Tree Topologies

2.1 An application of the birthday problem

In empirical phylogenomics, many gene trees on the same set of taxa (barring missing taxa) are estimated from multiple loci. Also, the relationships between the gene trees and species trees have been vastly studied (Pamilo and Nei, 1988; Takahata, 1989; Maddison, 1997; Rosenberg, 2002), but few studies have been done on the relationships among the gene trees given a species tree. It is frequently observed that due to certain evolutionary processes, species and gene trees are topologically discordant. These evolutionary processes that disrupt the equivalence of the gene and species tree topologies also lead to the incongruence of gene tree topologies given a species tree. Sometimes, every estimated gene tree has a unique topology (Salichos and Rokas, 2013). Also, this is reported to indicate that a data set has a large amount of heterogeneity and suggestions that processes such as incomplete lineage sorting (ILS) due to short branches in the species tree and other biological processes are likely the cause. However, the relative roles of these processes or systematic gene tree estimation error in causing this variability are not yet well understood. Investigation of factors of the gene tree heterogeneity is the key focus of this study.

Moreover, even for moderate numbers of species, the number of possible gene tree topologies is astronomically large. It is not clear when it should be surprising that every gene tree is unique. For example, for 4-taxon species, there are 15 possible rooted binary gene trees, while there are 43,459,425 possible rooted binary gene trees for 10-taxon species. In general, for any n-taxon species the number of possible rooted (or unranked) binary gene trees is (Felsenstein, 2004):

$$T_n = \frac{(2n-3)!}{2^{n-2}(n-2)!} \tag{2.1}$$

and the number of possible ranked gene trees is:

$$R_n = \frac{n!\,(n-1)!}{2^{n-1}} \tag{2.2}$$

Further, a linear relationship between R_n and T_n is:

$$R_n = \frac{n!\,(n-1)!\,(n-2)!\,T_n}{2(2n-3)!} = \frac{n(n-1)^2[(n-2)!\,]^3}{2(2n-3)!} \tag{2.3}$$

The probability that each gene tree topology is unique is likely to increase with the number of taxa and some evolutionary processes like incomplete lineage sorting due to short branches in the species tree (Degnan and Salter, 2005). Furthermore, short alignments and mutation rates that are either too low or too high can increase the uncertainty in the gene trees, therefore leading to more variability in the estimated gene trees than occurs in the actual gene trees. Also, one can expect that the gene trees that deviated much from the species tree should have a high probability of incongruence. Increasing the number of loci, on the other hand, increases the chance that at least two gene trees share the same topology, which decreases the probability that all gene tree topologies are unique.

In addition to the question of uniqueness, we might be more generally interested in a species tree, and how variable gene tree distributions are. This can be measured by the expected number and variance of the number of different gene trees when they are not all unique and the average distance between pairs of gene trees or between the gene tree and species tree. Understanding the distributions of these incongruent and discordant measures, we can better understand whether the level of the variability in an empirical data set is unusual. It will also help to know how alleles evolve and sort within species.

2.2 Birthday Problem and Probability bounds

2.2.1 Uniform Case

We can think of the number of unique gene tree topologies sampled from a species tree as a particular case of the birthday problem. In a classical birthday problem, indistinguishable balls are thrown independently into a fixed number of the distinguishable boxes, with each ball having probability p_i of falling into the i^{th} box. The frequencies $(p_i, i = 1, 2...)$ are assumed to be positive with $\sum_i p_i = 1$. For the finite case, n balls are thrown into m boxes, the number of balls inside the boxes is captured by $b_n = (b_{n,1}, \ldots, b_{n,m})$, where $b_{n,i}$ is the number of balls that fall into box i after throwing the last ball. The vector b_n and other variables that depend on b_n such as the number of nonempty boxes $(\#\{i : b_{n,i} > 0\})$, and number of the boxes that contain a certain number of ball(s) (say, $1, 2, ..., n$) have been extensively studied both in finite and infinite cases of balls and boxes (Gnedin and Yakubovich, 2007). The most commonly studied birthday problem is finding the probability of shared birthdays in a random selection of n people. Here one is interested in the probability of the event $b_{n,i} > 1$ ($i = 1, ..., 365$) with the uniform assumption that each day of the year is equally probable. For example, the probability that each of the n individuals randomly selected has a unique birthday is

$$p(\text{all unique}) = 1 \cdot \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n-1}{365}\right)$$

assuming that (i) there are 365 days in a year (ignoring leap years), (ii) each birthday is equally likely, and (iii) all individuals are independent. Of course, by the pigeonhole principle, $p(\text{not all unique}) = 1$ when $n > 365$. Famously, the probability that not all birthdays are unique is a little above 0.5 when $n = 23$ and about 0.9 when $n = 42$. It is often surprising and sometimes counter-intuitive how few people are needed to get a high probability of coincidence of birthdays. In reality, birthdays are not equiprobable, but the equiprobable assumption makes it less likely for all birthdays to be unique. Thus, $n = 23$ provides a lowest upper bound for the number of individuals needed to have at least 50% probability that all are not unique.

In general, we think of the gene tree topologies as boxes and the observed gene trees in a sample as balls. Suppose there are k topologies (boxes), and n_g gene trees (balls) are distributed to the k topologies ($n_g < k$). Then the probability that each gene tree has a separate topology is

$$1 \cdot \left(1 - \frac{1}{k}\right)\left(1 - \frac{2}{k}\right) \cdots \left(1 - \frac{n_g - 1}{k}\right)$$

If the probability of each topology $p = \frac{1}{k}$, then the probability that all is unique can written as

$$P(\text{all unique}) = \prod_{i=1}^{n_g - 1} (1 - ip) \tag{2.4}$$

A difficulty is that gene tree topologies, like birthdays, are not equiprobable. However, the number of possible rooted or ranked binary gene trees is known and constant (equations (2.1) and (2.2)). But the frequencies of the gene tree topologies in samples are random and are likely to depend on: (i) some biological and environmental factors such as speciation, extinction, mutation, incomplete lineage sorting, and recombination; and (ii) structural and sampling factors such as the number of species and number of loci sampled. Considering only the structural factors and assuming the frequencies of the gene trees topologies are constant, then for 5-taxon species, $T_5 = 105$ and thus, for $n_g = 16$, we have the probability of no repetition of gene tree topology in our sample is less than 0.5, that is, the probability that at least two gene trees have a common topology is greater than 0.5. Also, for 10-taxon species tree, $T_{10} = 34,459,425$ and for $n_g = 7,000$, the probability of no repetition of gene tree topology in our sample is less than 0.5. Since the species tree and gene tree distributions depend heavily on biological and environmental factors, we use two approaches for dealing with this problem: (i) approximate bounds on the probability and (ii) driving expected values and variances for the number of distinct topologies. However, a simulation is necessary but limited to a few choices of parameters. We will discuss this in detail in Chapter 3.

To bound the probability from above, we instead consider the probability that all ranked gene trees (also called labeled histories) are unique. That is, we distinguish two trees that are topologically equivalent if they have a different order of coalescences. In this case, as the internal branches of the species tree approach length 0, the ranked gene tree probabilities approach being equiprobable, which will lead to larger possibilities that all ranked gene tree topologies are unique.

Let R_n be the number of possible rooted binary gene tree topologies for a given species tree. Let n_g be the sample size of gene trees from a given species tree. Holding other biological factors constant, in the limit, as the internal branch lengths of the species tree approach 0, the probability of no matching gene tree topologies is

$$P(\text{all unique}) = \prod_{i=1}^{n_g} \left(1 - \frac{i-1}{R_n}\right) = \frac{R_n!}{(R_n - n_g)!\,(R_n)^{n_g}} \tag{2.5}$$

Letting $p = \frac{1}{R_n}$ in (2.5), we obtain

$$P(\text{all unique}) = \prod_{i=1}^{n_g} (1 - (i-1)p)$$

Now, using the inequality $1 - x \leq e^{-x}$, we obtain that

$$P(\text{all unique}) \leq \prod_{i=1}^{n_g} e^{-(i-1)p} \tag{2.6}$$

Substituting back $p = \frac{1}{R_n}$ in (4), we have

$$P(\text{all unique}) \leq \prod_{i=1}^{n_g} e^{-\frac{i-1}{R_n}} = e^{-\sum_{i=1}^{n_g} \frac{i-1}{R_n}} = e^{-\frac{1}{R_n}\sum_{i=0}^{n_g-1} i} = e^{-\frac{(n_g-1)n_g}{2R_n}}$$

So, an upper bound for the probability that every topology of gene trees in a sample of size n_g from a given species tree is unique is $e^{-\frac{(n_g-1)n_g}{2R_n}}$. This bound goes to zero as $n_g \to \infty$ and tends to $e^{-(\frac{R_n-1}{2})}$ as $n_g \to R_n$. Also, it increases to 1 as the number of species increases since R_n increases faster than exponentially with the number of species, n. So, as the sample size increases, the probability that every gene tree topology sampled is unique decreases but increases as the number of species increases.

Number of species	7	8	9	10	11	12
Probability of Uniqueness	0.012	0.854	0.996	1.000	1.000	1.000

Table 2.1: The probabilities of all-unique with sample size of 1000 for $7, 8, 9, 10, 11$, and 12 species.

Also, using Taylor expansion, minorization and majorization processes we obtain a lower and an upper bounds for the probability of uniqueness of gene trees topologies given a species tree, assuming that some biological factors are constant.

Theorem 2.2.1. *Let n_g be the number of gene trees sampled from T_n, possible rooted binary gene trees for a given species tree of n taxa with $n_g < R_n$ (number* *of ranked rooted binary trees). Assuming that the internal branches of the species tree approach length 0 and the probabilities of gene tree topologies are equiprobable and other biological factors are constant. Then*

$$e^{-\frac{2n_g^2(2n-4)!T_n}{(n-1)^2[(n-2)!]^3T_n-2n_g(2n-3)!}} \leq p(\mathit{all\ unique}) \leq e^{-\frac{(n_g-1)n_g}{2T_n}} \tag{2.7}$$

Before the proof of Theorem 2.2.1, we state the following Lemma and its proof is on Appendix A.

Lemma 2.2.2. *For any* $k \geq 0$, $\sum_{i=1}^{m-1} i^k \leq \frac{m^{k+1}}{k+1}$

Proof of Theorem 2.2.1

The probability of uniqueness from (2.5) is:

$$p(\text{all unique}) = \prod_{g=1}^{n_g}\left(1 - \frac{g-1}{R_n}\right)$$

Now, using Taylor series expansion (as in $-\ln(1-t) = \sum_{i=1}^{\infty}\frac{t^i}{i}$), we get

$$-\ln[p(\text{all unique})] = \sum_{g=1}^{n_g}\sum_{i=1}^{\infty}\frac{(g-1)^i}{i(R_n)^i}$$

So,

$$\begin{aligned}
-\ln[p(\text{all unique})] &= \sum_{g=1}^{n_g}\sum_{i=1}^{\infty}\frac{(g-1)^i}{i(R_n)^i} \\
&= \sum_{g=1}^{n_g}\left(\frac{g-1}{R_n}+\frac{(g-1)^2}{2R_n^2}+\frac{(g-1)^3}{3R_n^3}+\frac{(g-1)^4}{4R_n^4}+...\right) \\
&= \frac{1}{R_n}\sum_{g=1}^{n_g}(g-1)+\frac{1}{2R_n^2}\sum_{g=1}^{n_g}(g-1)^2+\frac{1}{3R_n^3}\sum_{g=1}^{n_g}(g-1)^3+... \\
&= \frac{(n_g-1)n_g}{2R_n}+\frac{(n_g-1)n_g(2n_g-1)}{12R_n^2}+...
\end{aligned}$$

Dropping the terms after the first term we get,

$$-\ln[p(\text{all unique})] \geq \frac{(n_g-1)n_g}{2R_n}$$

and

$$p(\text{all unique}) \leq e^{-\frac{(n_g-1)n_g}{2R_n}} \tag{2.8}$$

Also, for the lower bound, recall that,

$$\begin{aligned}
-\ln[p(\text{all unique})] &= \frac{1}{R_n}\sum_{g=1}^{n_g}(g-1)+\frac{1}{2R_n^2}\sum_{g=1}^{n_g}(g-1)^2+\frac{1}{3R_n^3}\sum_{g=1}^{n_g}(g-1)^3+... \\
&= \frac{1}{R_n}\sum_{g=1}^{n_g-1}g+\frac{1}{2R_n^2}\sum_{g=1}^{n_g-1}g^2+\frac{1}{3R_n^3}\sum_{g=1}^{n_g-1}g^3+\frac{1}{4R_n^4}\sum_{g=1}^{n_g-1}g^4+...
\end{aligned}$$

Using Lemma 2.2.2, we get

$$-\ln[p(\text{all unique})] \leq \frac{n_g^2}{2R_n}+\frac{n_g^3}{6R_n^2}+\frac{n_g^4}{12R_n^3}+\frac{n_g^5}{20R_n^4}+...$$

Replacing the constant factors in the denominators with 1 in the above series and factoring out n_g, we get:

$$-\ln[p(\text{all unique})] \;\leq\; n_g\left(\frac{n_g}{R_n}+\frac{n_g^2}{R_n^2}+\frac{n_g^3}{R_n^3}+\frac{n_g^4}{R_n^4}+...\right)$$

Since $n_g < R_n$ and the series in the bracket is geometric we have,

$$-\ln[p(\text{all unique})] \leq \frac{n_g^2}{R_n - n_g} \tag{2.9}$$

Substituting (2.3) in (2.9) for R_n, we have

$$\begin{aligned}
-\ln[p(\text{all unique})] \;&\leq\; \frac{n_g^2}{\frac{n(n-1)^2[(n-2)!]^3T_n}{2(2n-3)!} - n_g}\\
&=\; \frac{2n_g^2(2n-3)!}{n(n-1)^2[(n-2)!\,]^3T_n - 2n_g(2n-3)!}\\
&\leq\; \frac{2n_g^2(2n-3)(2n-4)!\cdot(\frac{T_n}{(2n-3)})}{n(n-1)^2[(n-2)!\,]^3T_n - 2n_g(2n-3)!}\\
&=\; \frac{2n_g^2(2n-4)!\,T_n}{n(n-1)^2[(n-2)!\,]^3T_n - 2n_g(2n-3)!}
\end{aligned}$$

Therefore,

$$p(\text{all unique}) \geq e^{-\frac{2n_g^2(2n-4)!T_n}{n(n-1)^2[(n-2)!]^3T_n - 2n_g(2n-3)!}} \tag{2.10}$$

Combining (2.8) and (2.10) proves Theorem 2.2.1.

n_g	6-taxon	7-taxon	8-taxon	9-taxon	10-taxon
1	0.981	0.992	0.997	0.999	1.000
2	0.925	0.967	0.987	0.995	0.998
3	0.839	0.928	0.971	0.989	0.996
4	0.732	0.875	0.949	0.981	0.994
5	0.615	0.812	0.921	0.971	0.990
6	0.496	0.741	0.889	0.958	0.986
7	0.385	0.665	0.852	0.944	0.981
8	0.288	0.587	0.811	0.927	0.975
9	0.206	0.509	0.767	0.909	0.969
10	0.143	0.435	0.721	0.888	0.961
11	0.095	0.365	0.673	0.867	0.953
12	0.060	0.301	0.624	0.843	0.945
13	0.037	0.245	0.575	0.819	0.936
14	0.022	0.195	0.526	0.793	0.926
15	0.012	0.153	0.479	0.766	0.915
16	0.007	0.118	0.433	0.739	0.904
17	0.004	0.090	0.388	0.711	0.892
18	0.002	0.067	0.346	0.682	0.880
19	0.001	0.049	0.307	0.653	0.867
20	0.0004	0.036	0.270	0.623	0.854
25	0.000	0.005	0.129	0.478	0.782
30	0.000	0.001	0.053	0.345	0.701
35	0.000	0.00004	0.018	0.235	0.617
40	0.000	0.0000	0.005	0.151	0.532
45	0.000	0.0000	0.001	0.091	0.450
50	0.000	0.000	0.0003	0.052	0.373

Table 2.2: Lower bounds for the probability of uniqueness for $6, 7, 8, 9$ and 10 taxon species trees.

Theoretically, holding other biological factors constant, the probability that all tree topologies are unique depends on the number of the tips of the species tree and number of loci sampled.

2.3 Expected Value and Variance of the Number of Unranked Unique Gene Tree Topologies

Suppose we have **3-taxon** species trees, then $T_n = 3$. Let p_1, p_2, p_3 be the probabilities of the topologies. The p_i's, $i = 1, 2, 3$, are not necessarily the same. Now, for a sample size of 2, we have

$$p(\text{uniqueness}) = p_1p_2 + p_2p_1 + p_1p_3 + p_3p_1 + p_2p_3 + p_3p_2 = 2(p_1p_2 + p_1p_3 + p_2p_3)$$

Also, for sample size of 3, we have

$$p(\text{uniqueness}) = 3!\, p_1p_2p_3$$

For **4-taxon** species trees, the probabilities are $p_1, p_2, ..., p_{15}$. For a sample of size 2, the probability that the gene trees have distinct topologies is

$$p(\text{uniqueness}) = 2(p_1p_2 + p_1p_3 + p_1p_4 + ... + p_{14}p_{15}) = 2\sum_{i<j} p_ip_j$$

For a sample of size 3, we have that for $i, j, k \in \{1, 2, ..., 15\}$,

$$p(\text{uniqueness}) = 3! \sum_{i<j<k} p_ip_jp_k$$

In general, for n-**taxon** species trees, and a sample of size n_g, we have

$$p(\text{uniqueness}) = n_g! \sum_{i_1<i_2<...<i_{n_g}} p_{i_1}p_{i_2}...p_{i_{n_g}}$$

Let

$$X_{i,j} = \begin{cases} 1 & \text{if jth locus is gene tree i} \\ \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

Recall that T_n is the possible number of bifurcating unranked gene tree topologies.

Now, the number of times topology i occurs in a sample of size n_g is:

$$X_i = \sum_{j=1}^{n_g} X_{i,j} \tag{2.12}$$

where $i = 1, ..., T_n$, $j = 1, ..., n_g$

and

$$\sum_{i=1}^{T_n} X_i = \sum_{i=1}^{T_n} \sum_{j=1}^{n_g} X_{i,j} = n_g$$

Also, let Y_i be an indicator that topology i occurs at least once in the sample:

$$Y_i = \begin{cases} 1 & \text{if } X_i > 0 \\ \\ 0 & \text{if } X_i = 0 \end{cases} \tag{2.13}$$

Then the number of distinct topologies is

$$Y = \sum_{i=1}^{T_n} Y_i$$

Here, $X_{i,j}$ is a Bernoulli random variable with success probability p_i and X_i's are Binomial random variables with parameters p_i and m for each i.

Further,

$$\begin{aligned}
E[Y_i] &= 1 \cdot P(X_i > 0) + 0 \cdot P(X_i = 0) \\
&= P(X_i > 0) = 1 - P(X_i = 0) \\
&= 1 - (1 - p_i)^{n_g} \\
E[Y_iY_j] &= P(Y_iY_j = 1) = P(X_i > 0, X_j > 0) = 1 - P(X_i = 0 \text{ or } X_j = 0)
\end{aligned}$$

and the probability that $X_i = 0$ or $X_j = 0$ is:

$$\begin{aligned}
P(X_i = 0 \text{ or } X_j = 0) &= P(X_i = 0) + P(X_j = 0) - P(X_i = 0 \text{ and } X_j = 0) \\
&= (1 - p_i)^{n_g} + (1 - p_j)^{n_g} - (1 - p_i - p_j)^{n_g}
\end{aligned}$$

Then

$$E[Y_iY_j] = 1 - (1 - p_i)^{n_g} - (1 - p_j)^{n_g} + (1 - p_i - p_j)^{n_g}$$

and

$$\begin{aligned}
Var[Y_i] &= [1 - (1 - p_i)^{n_g}] \cdot [1 - (1 - (1 - p_i)^{n_g})] \\
&= (1 - p_i)^{n_g} - (1 - p_i)^{2n_g}
\end{aligned}$$

Also, the covariance of Y_i and Y_j, $i \neq j$, is:

$$\begin{aligned} Cov(Y_i, Y_j) &= E[Y_iY_j] - E[Y_i]E[Y_j] \\ &= 1 - (1-p_i)^{n_g} - (1-p_j)^{n_g} + (1-p_i-p_j)^{n_g} \\ &- (1-(1-p_i)^{n_g})\,(1-(1-p_j)^{n_g}) \\ &= (1-p_i-p_j)^{n_g} - (1-p_i)^{n_g}(1-p_j)^{n_g} \end{aligned}$$

2.3.1 The expected number and variance of distinct topologies

The expected number of distinct topologies is:

$$\begin{aligned} E[Y] &= \sum_{i=1}^{T_n} E[Y_i] = \sum_{i=1}^{T_n} [1-(1-p_i)^{n_g}] \\ &= T_n - \sum_{i=1}^{T_n} (1-p_i)^{n_g} \end{aligned}$$

and the variance is:

$$\begin{aligned} Var[Y] &= \sum_{i=1}^{T_n} Var[Y_i] + 2 \sum_{1\le i<j\le T_n} Cov(Y_i, Y_j) \\ &= \sum_{i=1}^{T_n} [(1-p_i)^{n_g} - (1-p_i)^{2n_g}] \\ &+ 2 \sum_{1\le i<j\le T_n} [(1-p_i-p_j)^{n_g} - (1-p_i)^{n_g}(1-p_j)^{n_g}] \end{aligned}$$

Then, as $n_g \to \infty$, $E[Y]$ and $Var(Y)$ converge to T_n and 0, respectively. Table 2.4 depicts the behaviour of expected number of distinct topologies as the sample size increases.

2.3.2 Uniform Case

For the equal probable case -that is: $p_i = \frac{1}{T_n}$ for each i. We have

$$\begin{aligned} E[Y] &= T_n - T_n\left(1 - \frac{1}{T_n}\right)^{n_g} \\ &= T_n\left[1 - (1 - \frac{1}{T_n})^{n_g}\right] \end{aligned}$$

and

$$\begin{aligned} Var[Y] = T_n &\left[(1 - \frac{1}{T_n})^{n_g} - (1 - \frac{1}{T_n})^{2n_g}\right] \\ &+ 2\left[T_n(1 - 2\frac{1}{T_n})^{n_g} - (1 - \frac{1}{T_n})^{2n_g}\right] \end{aligned}$$

where $1 \leq n_g \leq T_n$.

For example, we enumerate the all 15 possible topologies of 4-taxon species tree with their probabilities on Table 2.3, and behavior of the expected number of distinct gene tree topologies as the sample size (loci) increases 2.4.

	Topology	Probability
1	(((A,D),B),C);	0.001
2	((A,(B,D)),C);	0.001
3	(((A,B),D),C);	0.098
4	((A,B),(C,D));	0.099
5	(((A,B),C),D);	0.556
6	(((A,D),C),B);	0.001
7	((A,(C,D)),B);	0.001
8	(((A,C),D),B);	0.021
9	((A,C),(B,D));	0.022
10	(((A,C),B),D);	0.079
11	(((B,D),C),A);	0.001
12	((B,(C,D)),A);	0.001
13	(((B,C),D),A);	0.021
14	((B,C),(A,D));	0.022
15	(((B,C),A),D);	0.079

Table 2.3: Topologies and their probabilities for 4-taxon species tree, (((A,B),C),D).

n_g / n	4-taxon	5-taxon	6-taxon	7-taxon	8-taxon
100	9.102	19.593	32.701	46.426	59.235
1,000	12.832	37.244	86.785	164.847	265.970
10,000	15.000	62.135	175.107	403.547	808.495
100,000	15.000	78.158	292.258	841.088	1963.048
1,000,000	15.000	86.350	402.687	1397.214	3986.560
100,000,000	15.000	105.000	641.611	2945.781	10864.422

Table 2.4: The sample sizes (loci) and the expected numbers of distinct gene tree topologies for $4-8$ taxon species trees. This indicates that the expected number of distinct topologies increases with the number of loci sampled.

2.3.3 Missing Topologies

Here we find the expected value and variance of the number of topologies that are not in a sample of n_g gene trees. Recall, T_n is the total number of possible gene tree topologies for a species tree of n taxa and Y is the number of distinct topologies. Then

$$M = T_n - Y$$

is the number of missing gene tree topologies. The expected number of missing topologies of is:

$$\begin{aligned} E[M] &= E[T_n - Y] = T_n - E[Y] \\ &= T_n - \sum_{i=1}^{T_n} E[Y_i] \\ &= T_n - (T_n - \sum_{i=1}^{T_n} (1-p_i)^{n_g}) \\ &= \sum_{i=1}^{T_n} (1-p_i)^{n_g} \end{aligned}$$

and

$$\begin{aligned} Var(M) &= Var(T_n - Y) = Var(Y) \\ &= \sum_{i=1}^{T_n} [(1-p_i)^{n_g} - (1-p_i)^{2n_g}] \\ &+ 2 \sum_{1 \le i < j \le T_n} [(1-p_i-p_j)^{n_g} - (1-p_i)^{n_g}(1-p_j)^{n_g}] \end{aligned}$$

Both $E[M]$ and $Var(M)$ go to zero as the number of loci approaches ∞ regardless of the number of taxa. Note that the convergence of these series is guaranteed since $T_n < \infty$. Thus, the probability of missing topologies decreases as the number of loci increases.

2.4 Non-Uniform Case

As mentioned in the previous section, the topologies of gene trees from a given species tree are not uniformly distributed. Similarly, birthdays are not uniformily distributed in a calendar year. For instance, some birthdays are pre-planned due to circumstances like C-sections, inductions, days of hospital operations, etc. Borja (2016) noted that there are more births in spring and summer in Europe and America; and more births between Tuesdays and Fridays. These suggest that most conceptions are around Summers and Christmas holidays, and most deliveries through C-sections and inductions are on weekdays. Any apparent departures from the uniformity assumption would increase the likelihood that a sample of fewer sizes (less than 23) will contain some pairs with common birthdays. The same scenario applies to gene tree topologies. Intuitively, the likelihood of more gene trees having the same topology increases on the case of the non-uniformity assumption as in the case of the birthday problem(s). However, Nunnikhoven (1992) showed that the differences between the solutions of non-uniformity and uniformity models are minimal for the birthday problem. So, the assumption of uniformity approximates reality well. The birthday problems have been vastly generalized in several ways (Klamkin and Newman, 1967; Flajolet et al., 1992; Nunnikhoven, 1992; Mase, 1992; Camarri and Pitman, 2000; DasGupta, 2005; Feller, 2008). In the case of tree topologies, unranked topologies do not have uniform probabilities even if all the branch lengths are zero. When some branches in the species tree have non-zero length then this makes the probabilities even further from uniform.

Gene trees' discordance with their parent species tree or incongruence among gene trees can be caused by some evolutionary events (Pamilo and Nei, 1988; Rosenberg, 2002; Degnan and Salter, 2005). Such evolutionary processes would affect the uniformity assumption of the frequencies of gene tree topologies. Let p_i be the probability that a gene tree has topology i and n_g be the sample size of gene trees from a given species tree with n taxa. From equations (2.1), (2.11) and (2.12),

$$(X_i, ..., X_{T_n} | \sum_{i=1}^{T_n} X_i = n_g) \sim Mult(n_g, p_1, ..., p_{T_n}) \tag{2.14}$$

where $p_1, ..., p_{T_n}$ are non-negative reals such that $0 < p_i < 1$ and $\sum_{i=1}^{T_n} p_i = 1$. We note that the collection of counts $(X_1, ..., X_{T_n})$ is a multinomial distribution.

The probability of uniqueness (Mallows, 1968) is:

$$p(\text{all unique}) = p(X_1 \leq 1, ..., X_{T_n} \leq 1) \leq p(X_1 \leq 1)...p(X_{T_n} \leq 1) \tag{2.15}$$

Theorem 2.4.1. *Let $p_1, p_2, ..., p_{T_n}$ be probabilities of gene tree topologies and p_i's are not necessarily equal. Let $p_{min} = \min_{1 \leq i \leq T_n}\{p_i\} = \frac{\alpha_n}{T_n}$ and $p_{max} = \max_{1 \leq i \leq T_n}\{p_i\} = \frac{\gamma_n}{T_n}$ with $0 < \alpha_n \leq \gamma_n \leq T_n$ for some real numbers α_n and γ_n that depend on the number of taxa of the species tree. Then for $n_g \leq T_n$,*

$$p(\textit{all unique}) \leq \exp\{-n_g[1-((1-\frac{\alpha_n}{T_n})^{n_g} + n_g\frac{\gamma_n}{T_n}(1-\frac{\alpha_n}{T_n})^{n_g-1})]\} \tag{2.16}$$

and for $n_g > T_n$,

$$p(\textit{all unique}) = 0 \tag{2.17}$$

Proof

From (2.15), we obtain

$$p(\text{all unique}) \leq \prod_{i=1}^{T_n} p(X_i \leq 1)$$

Also, for any real number x, $1 - x \leq e^{-x}$, so,

$$p(X_i \leq 1) = 1 - p(X_i > 1) \leq e^{-p(X_i \geq 1)}$$

for each i. Then

$$\begin{aligned}
p(\text{all unique}) &\leq \prod_{i=1}^{T_n} p(X_i \leq 1) = \prod_{i=1}^{T_n} (1 - p(X_i > 1) \\
&\leq \prod_{i=1}^{T_n} \exp\{-p(X_i > 1\} = \exp\{-\sum_{i=1}^{T_n} p(X_i > 1)\} \\
&\leq \exp\{-\sum_{i=1}^{n_g} p(X_i > 1)\} = \exp\{-\sum_{i=1}^{n_g} (1 - p(X_i \leq 1))\} \\
&= \exp\{-\sum_{i=1}^{n_g} (1 - [p(X_i = 0) + p(X_i = 1)])\} \\
&= \exp\{-\sum_{i=1}^{n_g} [1 - ((1-p_i)^{n_g} + n_g p_i (1-p_i)^{n_g-1})]\} \\
&= \exp\{-n_g + \sum_{i=1}^{n_g} [(1-p_i)^{n_g} + n_g p_i (1-p_i)^{n_g-1}]\} \\
&\leq \exp\{-n_g + \sum_{i=1}^{n_g} [(1-p_{min})^{n_g} + n_g p_{max} (1-p_{min})^{n_g-1}]\} \\
&= \exp\{-n_g + n_g[(1-p_{min})^{n_g} + n_g p_{max} (1-p_{min})^{n_g-1}]\} \\
&= \exp\{-n_g[1 - ((1-p_{min})^{n_g} + n_g p_{max} (1-p_{min})^{n_g-1})]\} \\
&= \exp\{-n_g[1 - ((1-\frac{\alpha_n}{T_n})^{n_g} + n_g \frac{\gamma_n}{T_n} (1-\frac{\alpha_n}{T_n})^{n_g-1})]\}
\end{aligned}$$

Also, by the pigeon hole principle, (2.17) is true.

Also, from (2.14) and according to Levin (1981), the probability of uniqueness is:

$$p(\text{all unique}) = p\left(X_1 \leq 1, ..., X_{T_n} \leq 1 | \sum_{i=1}^{T_n} X_i = n_g\right) \tag{2.18}$$

We note that based on the conditional probability (2.18), X_i's are not independent, but the application of Bayes' Theorem allows us to remove the condition that the sum is fixed. Then, applying Bayes' Theorem, we obtain:

$$p(\text{all unique}) = \frac{p\left(\sum_{i=1}^{T_n} X_i = n_g | X_i \leq 1,\ \forall\ i\right) p\left(X_i \leq 1,\ \forall\ i\right)}{p\left(\sum_{i=1}^{T_n} X_i = n_g\right)} \tag{2.19}$$

But by construction, $p(\sum_{i=1}^{T_n} X_i = n_g) = 1$ since $\sum_{i=1}^{T_n} X_i = n_g$ always. So,

$$p(\text{all unique}) = p\left(\sum_{i=1}^{T_n} X_i = n_g | X_i \leq 1,\ \forall\ i\right) p\left(X_i \leq 1,\ \forall\ i\right) \tag{2.20}$$

Without loss of generality, p_i is small for each i for a species tree with large number of taxa since T_n grows faster than exponential, Thus, X_i approximately has a Poisson distribution with parameter $\lambda = n_g p_i$. Let $H = \sum_{i=1}^{T_n} X_i$, then H is a sum of T_n independent right truncated Poisson random variables.

2.5 Discussion and Conclusions

In this chapter, we investigated the contributions of the number of tips of a species tree and the gene trees' sample size to the gene trees' heterogeneity. Assuming that the probability of gene tree topology is uniform and holding speciation, extinction, and mutation rates constant, we derived a bound for the probability that every gene tree has a unique topology given the sample size and the number of tips of the species tree. Based on the bound, the probability of uniqueness increase with the number of the tips of the species tree and decreases as the number of loci (sample size) increases. For example, for the 10-taxon species tree, 50 loci, the lower and upper bounds are 0.373 and 0.999, respectively. For 100 loci, the lower and upper are 0.0194 and 0.99986, respectively. Also, for a small number of tips of the species tree, the bound is relatively tight and widening as the number of the tips increases. We also derived the expected number and the variance of the number of distinct gene tree topologies for any given number of tips of the species tree and the number of loci. We found that the expected number and variance of the number of distinct gene trees converge to T_n (or R_n) (the number of possible gene tree topologies for unranked or ranked rooted trees) and zero, respectively, for any species tree as the number of loci tends to infinity.

Further, if the uniformity assumption is violated, the probability that all gene trees sampled have unique topology decrease, and is likely to decrease faster as the branch lengths of the species get shorter. We investigate the probabilities of uniqueness of gene tree topologies from larger species trees without any biological constant in detail with simulations in the next chapter.

Chapter 3

Incongruence of Gene Trees

Multilocus phylogenetic studies often estimate separate gene trees for each locus. Species trees can be inferred either directly from the sequence data or using a two-stage approach from the gene trees estimated from different loci (Liu et al., 2009; Rannala et al., 2020). There is often considerable heterogeneity in the gene tree topologies, with many distinct topologies being observed in many data sets. In some cases, every gene tree topology is distinct, even with over 1000 loci (Salichos and Rokas, 2013).

An interesting question is how much variation in the gene trees should be expected to be observed. There are many sources of this variability, some of which are biological, and some having to do more with statistical issues. In particular, horizontal gene transfer (HGT), gene duplication/loss (GDL), and the multispecies coalescent (MSC) all lead to variability in the gene trees, even if the gene trees are known perfectly (Maddison, 1997). Under the MSC, short internal branches in the species tree typically predict higher levels of discordance (i.e., topological discrepancies) between gene trees and the species tree, and also higher levels of incongruence—variation in gene tree topologies. The heterogeneity of the gene trees reconstructed from genomes given a species trees can be caused by either methodological processes or systematic errors. However, the relative roles of these processes or systematic error in causing this variability are not yet well understood. This investigation also examines the additional heterogeneity added when gene trees are estimated under the MSC.

There are several possible ways of characterizing the heterogeneity in a set of gene trees. For example, one can examine conflicting rooted triple relationships by counting the number of each of three possible triplets are supported in the collection of gene trees (Cranston et al., 2009). Another approach is to measure the average Robinson-Foulds (RF) distance between the gene trees and the species tree (Mirarab et al., 2014). This has the disadvantage that the species tree is often unknown. A proxy to this quantity, which we examine in this paper, is the average pairwise RF distance between all pairs of gene trees. In our simulations, this quantity is highly correlated with the gene tree-species tree average RF distance.

Another measure of heterogeneity is the number of distinct gene tree topologies in a sample. in particular, in some data sets every estimated gene tree topology has a distinct topology. This occurs, for example, in Salichos and Rokas (2013), which included 23 species of yeast using 1070 loci, each with a unique gene tree topology that did not match the estimated species treee. The authors point to this result as evidence of incomplete lineage sorting and a high level of gene tree discordance. On the other hand, for this many taxa, there are over 5×10^{26} possible rooted, binary topologies (**?**). For this vast number of possibilities, is it surprising that every gene tree has a unique topology? For purely random trees generated under a birth-death process, it would not be surprising for 1070 trees of this size to be unique. However, if gene trees evolved within a species tree, they should have much more in common than purely random trees. In this case, it is not clear whether this level of gene tree heterogeneity should be surprising.

We examine this question using simulation with species trees generated by a birth-death process and gene trees generated from each species tree using the MSC. We examine both true gene trees and gene trees estimated from DNA sequence alignments to see the effect that sampling error has on increasing gene tree heterogeneity. For the birth-death process, higher levels of speciation predict shorter internal branches in the species tree, making heterogeneity in the gene trees more likely. We use this framework to examine the effect of speciation rates, mutation rates, and alignment length on gene tree heterogeneity measured both by RF distances (to the species tree and pairwise) and by the number of distinct gene tree topologies. We find that results such as those observed by Salichos and Rokas (2013) are not necessarily surprising for moderately high speciation rates but can be surprising for lower speciation rates.

3.1 Simulation Design

We examine gene tree heterogeneity as a function of the number of taxa (n), the speciation rate (λ), extinction rate (μ), population mutation rate (θ), and sample size (n_g) (i.e., number of loci).

To investigate the effect of these factors on gene tree variability, we generated random species trees under a constant-rate birth-death model with speciation rate $\lambda = (0.1, 0.2, 0.5, 1.0)$, extinction rate $\mu = (0, 0.5\lambda)$, and number of taxa $n = (5, 10, 15, 20)$ with the R package TreeSim (Stadler, 2011a). We interpret the branch lengths for trees generated from TreeSim as being in coalescent units; i.e., a branch length of 1.0 means N generations where N is the effective diploid population size.

The range of $\lambda = 0.1$ to 1.0 allows the level of ILS to range from moderately low to fairly high, and is comparable to other simulation studies (Mirarab et al., 2014; Stadler et al., 2016a; Kim and Degnan, 2021). For each parameter combination, 100 species trees were simulated. An outgroup was then added to each species tree where the distance from the outgroup to the root was 10 coalescent units plus the height of the original tree. Then either $n_g = 500$ and 1000 gene trees were simulated for each species tree using hybrid-Lambda (Zhu et al., 2015). We note that for cases with estimated gene trees, the number of parameter combinations for $(n, \lambda, \mu, n_g, L, \theta)$ is therefore $4 \times 4 \times 2 \times 2 \times 2 \times 2 = 256$. For gene trees that were known rather than estimated, there were 64 parameter combinations.

For each gene tree, we simulated DNA sequences of length 500 and 1000 nucleotides (nt) of the tree with seq-gen (Rambaut and Grassly, 1997) with population mutation rate $\theta = 0.002$ or 0.01 under the $GTR{+}I{+}G$ model with 10% invariant sites, four rate categories, and base frequencies of $0.4, 0.1, 0.2, 0.3$ for A, C, G, and T, respectively. Then we reconstructed and estimated the gene trees with IQTree (Nguyen et al., 2014). To allow for among-site variation and compositional heterogeneity, we used the $GTR + G + I$ model in IQTree to reconstruct and estimate the gene trees. In addition to reconstructed gene trees, true gene trees taken directly from hybrid-Lambda were used as well.

We counted the number of unique gene trees in each sample with PRANC (Kim et al., 2020) using the option `-utopo`, which converts each Newick string into a string which uniquely represents the clades of each tree. Furthermore, we calculated the rooted Robinson-Foulds (RF) distance (Robinson and Foulds, 1981) among the gene trees (pairwise RF) after the outgroup is removed, and between gene trees and species tree (RF-ST) with Treedist in PHYLIP (Felsenstein, 1993). The probabilities of uniqueness of gene tree topologies were computed from these reconstructed gene trees for each species tree.

3.2 Simulation results

We examine the effects of the number of taxa, the speciation rate, the extinction rate, and whether or not gene trees were estimated or known. In addition, we compare the RF distances of gene trees to the species tree with pairwise RF distances between the gene trees without using the species treee.

3.2.1 *Number of Taxa*

Holding sample size and other factors constant, the number of unique gene tree topologies increases rapidly as the number of taxa increases (Figure 3.1). In particular, the number of distinct gene tree topologies often approaches the maximum possible value. For $n = 5$ taxa, only 105 topolgies are possible, but as the number of species increases, the number of distinct gene tree topologies approaches the sample size of 500. The probability of uniqueness also increases with the number of taxa, but does so more rapidly for $\mu = 0$ than for $\mu = .5\lambda$ (Figures 3.2 and 3.3). Both the proportion of unique gene trees (i.e., the number of distinct gene trees divided by the number of loci) and probability of uniqueness increase as the sample size decreases for the trees with higher numbers of taxa (Tables 3.1, and 3.2).

3.2.2 *Speciation Rate*

Increasing the speciation rate (λ) decreases the average branch length and makes more short intervals probable (Stadler and Steel, 2012b; Stadler et al., 2016b), leading to increased heterogeneity in the gene trees. For $\lambda \geq .5$, the probability of uniqueness rapidly increases and is very close to 1.0 for both sample sizes for species trees with 15 or more taxa (Figure 3.2). Consequently, it would not be surprising or unusual for all sampled gene trees to be distinct for $n \geq 15$, $\lambda \geq 0.5$, and $\mu = 0$. Furthermore, the average pairwise RF and RF-ST distances increase rapidly with λ for $\mu = 0$ (Tables 3.1 and 3.2). Thus, the simulations indicated that the speciation rate plays a strong role in the heterogeneity of the gene trees.

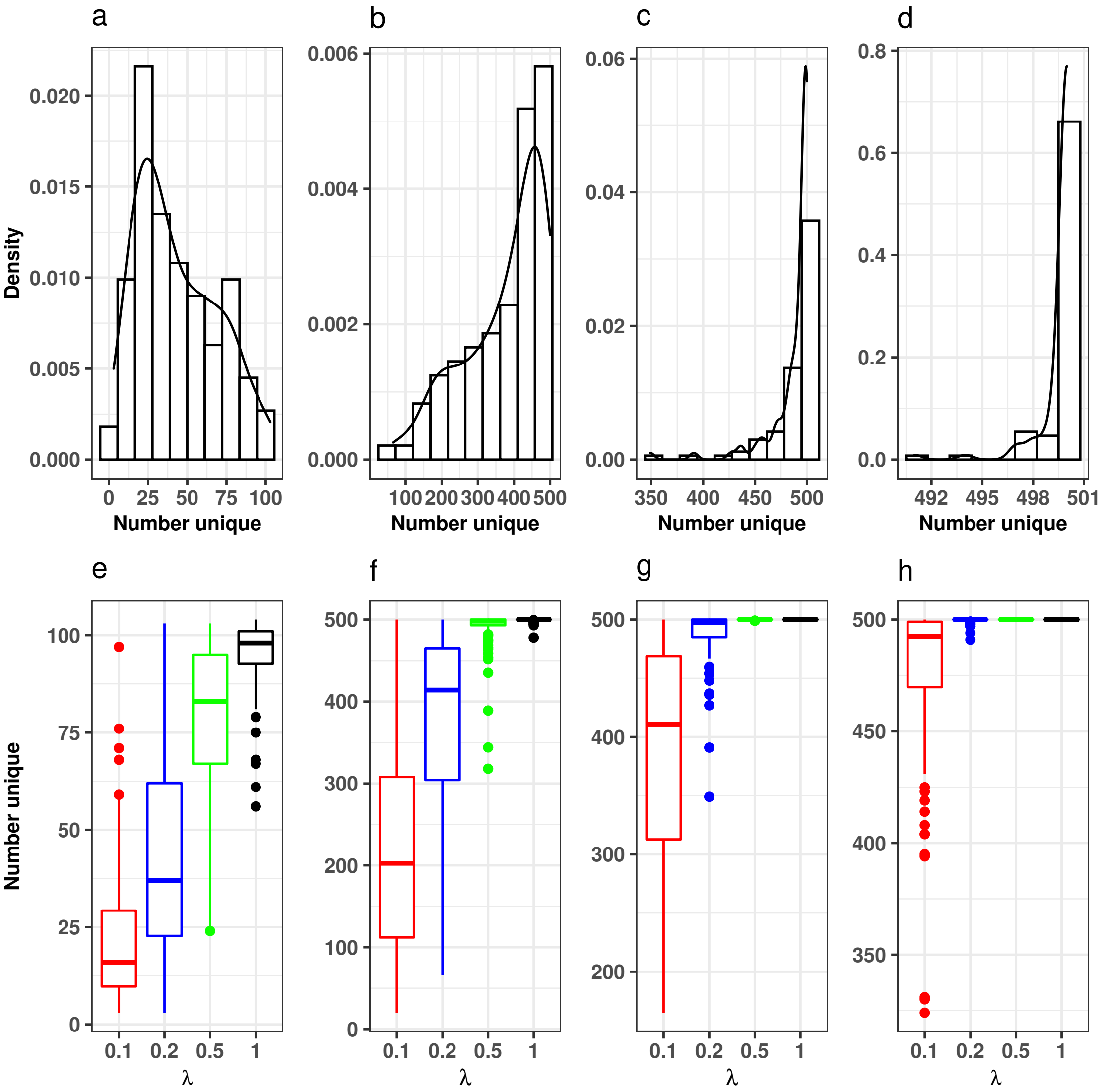

Figure 3.1: Plots (a)–(d) are distributional plots for the number of unique gene tree topologies when there are 5, 10, 15 and 20 species, respectively, for a sample size of 500 gene trees with $\mu = 0$, $\lambda = 0.2$, and $\theta = 0.002$. Plots (e)–(h) are boxplots for the number of unique gene trees as a function of the speciation rate λ when $\mu = 0$ in samples of size 500 gene trees for $n = 5, 10, 15$, and 20 species, respectively.

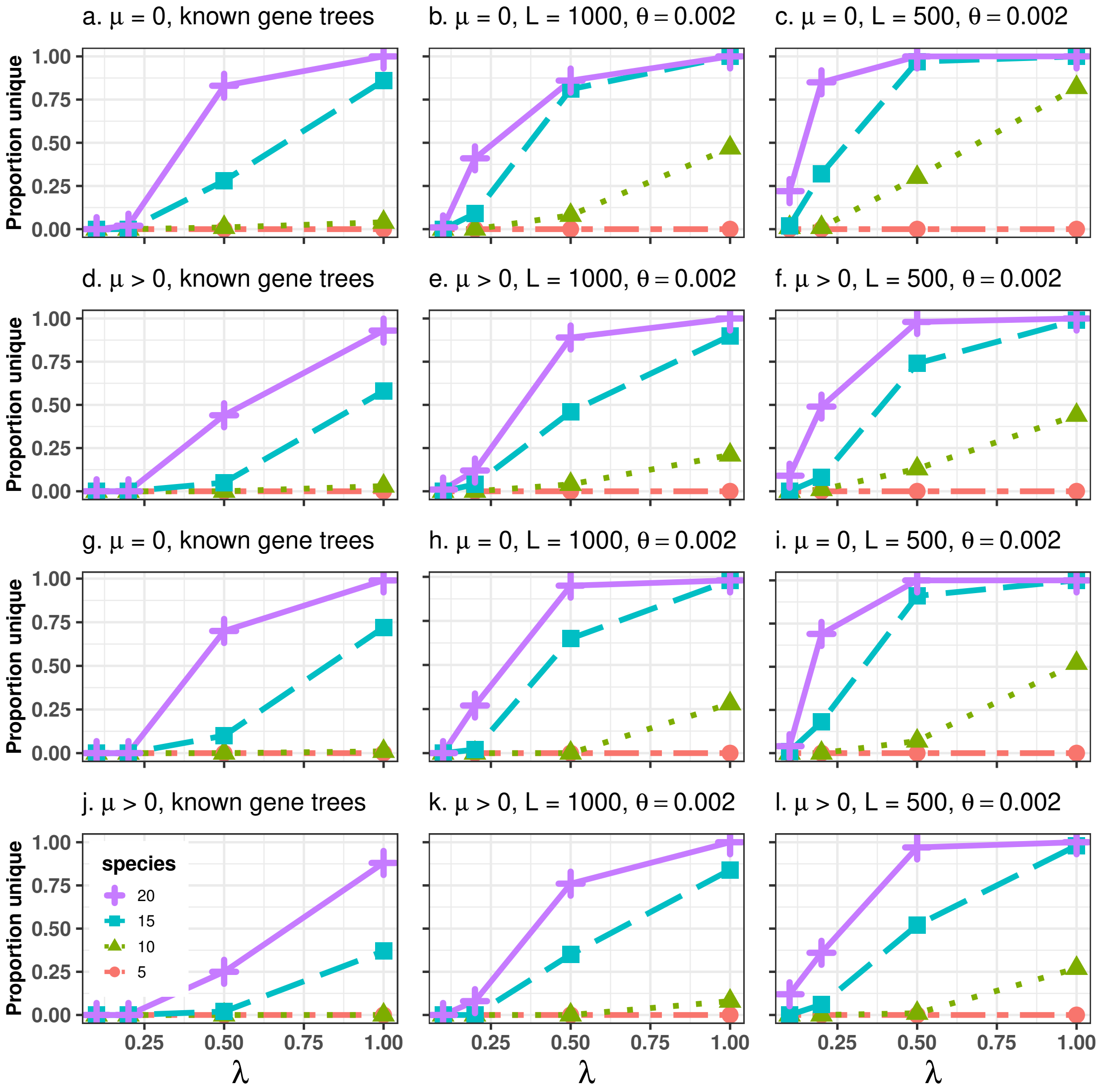

Figure 3.2: The estimated probability that every gene tree topology is unique for the sample size of 500. The legend in (j) applies for each subfigure. In (a)-(f), there were 500 loci per sample, and in (g)-(l), there were 1000 loci per sample. The first column used known gene trees, and for columns 2 and 3, $\theta = 0.002$ was used throughout to generate DNA sequences to estimate gene trees. The middle column used sequence lengths of 1000 nt for estimated gene trees, and the third column uses sequence lengths of 500 nt. For $\mu > 0$ (the second and fourth rows), $\mu = 0.5\lambda$.

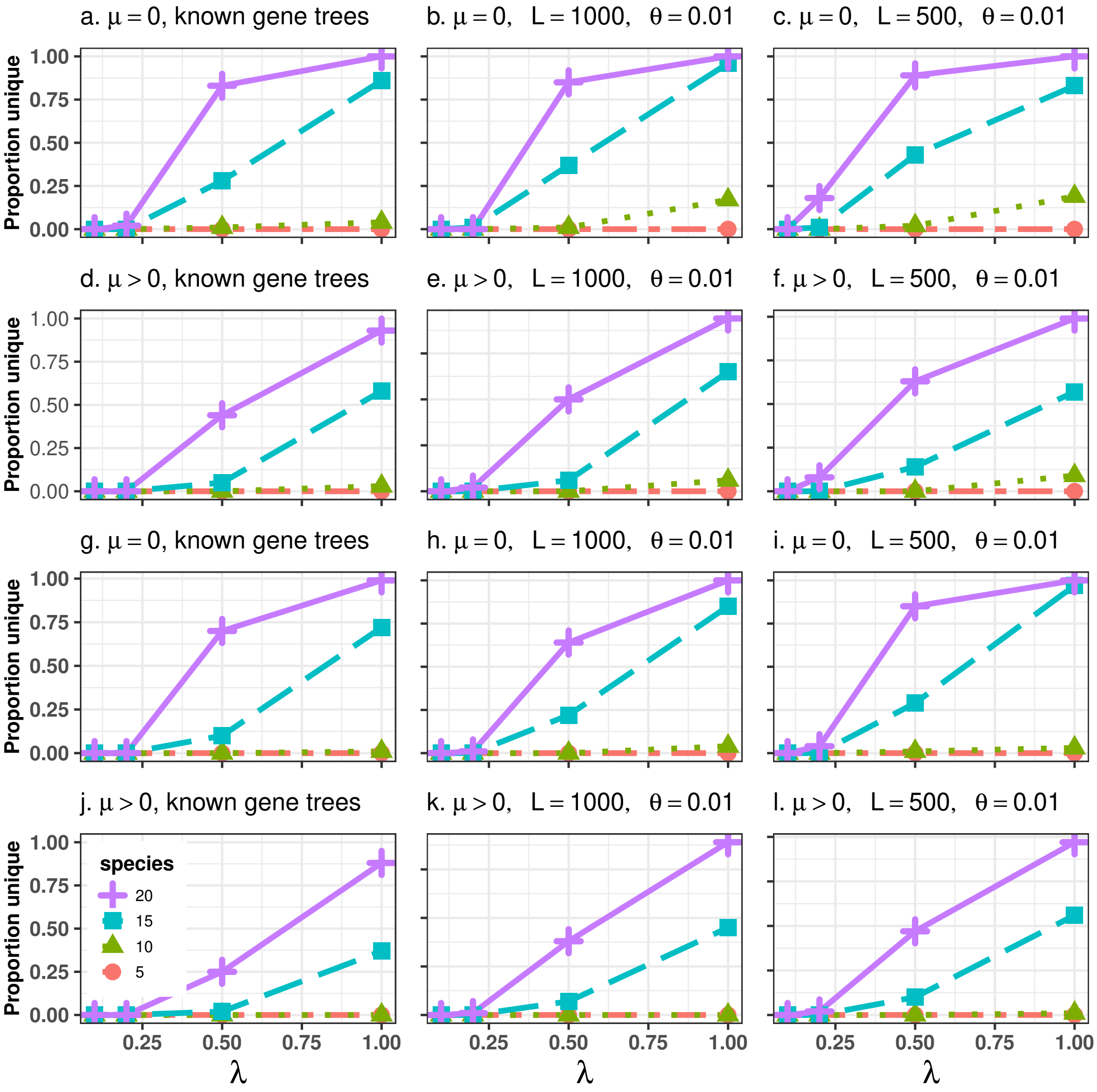

Figure 3.3: The estimated probability that every gene tree topology is unique for the sample size of 500. The legend in (j) applies for each subfigure. In (a)-(f), there were 500 loci per sample, and in (g)-(l), there were 1000 loci per sample. The first column used known gene trees, and for columns 2 and 3, $\theta = 0.01$ was used throughout to generate DNA sequences to estimate gene trees. The middle column used sequence lengths of 1000 nt for estimated gene trees, and the third column uses sequence lengths of 500 nt. For $\mu > 0$ (the second and fourth rows), $\mu = 0.5\lambda$.

3.2.3 *Extinction Rate*

Having a positive extinction rate affects the variability in the branch lengths and tends to make branches near the root of the tree longer. This results in more spe-

			λ			
θ	n	Sample size	0.1	0.2	0.5	1.0
NA	10	500	0.123	0.324	0.749	0.944
NA	10	1000	0.088	0.255	0.680	0.916
NA	15	500	0.278	0.641	0.975	1.000
NA	15	1000	0.217	0.566	0.960	0.999
NA	20	500	0.486	0.864	0.999	1.000
NA	20	1000	0.411	0.817	0.998	1.000
	Alignment length = $500nt$					
0.01	10	500	0.220	0.449	0.843	0.967
0.01	10	1000	0.161	0.363	0.785	0.958
0.01	15	500	0.499	0.798	0.992	0.994
0.01	15	1000	0.421	0.734	0.987	1.000
0.01	20	500	0.757	0.957	1.000	1.000
0.01	20	1000	0.693	0.934	1.000	1.000
0.002	10	500	0.422	0.752	0.978	0.999
0.002	10	1000	0.333	0.678	0.966	0.998
0.002	15	500	0.773	0.975	1.000	1.000
0.002	15	1,000	0.704	0.960	1.000	1.000
0.002	20	500	0.952	0.999	1.000	1.000
0.002	20	1,000	0.785	0.999	1.000	1.000
	Alignment length = $1000nt$					
0.01	10	500	0.183	0.414	0.817	0.967
0.01	10	1000	0.134	0.334	0.756	0.946
0.01	15	500	0.379	0.731	0.986	1.000
0.01	15	1000	0.308	0.660	0.978	1.000
0.01	20	500	0.623	0.763	0.999	1.000
0.01	20	1000	0.516	0.883	0.956	1.000
0.002	10	500	0.303	0.613	0.936	0.994
0.002	10	1000	0.236	0.531	0.902	0.989
0.002	15	500	0.572	0.903	0.999	1.000
0.002	15	1000	0.494	0.865	0.998	1.000
0.002	20	500	0.801	0.986	0.999	1.000
0.002	20	1000	0.741	0.976	1.000	1.000

Table 3.1: The average proportion of distinct gene trees (number of distinct gene trees divided by number of loci) given species trees of 15 and 20 taxa for the sample sizes of 500 and 1000 loci. The entry NA in the coluumn for θ means that known gene trees were used rather then estimated from DNA sequences. The extinction parameter μ was 0 throughout.

			λ			
θ	n	Sample size	0.1	0.2	0.5	1.0
NA	10	500	0.076	0.120	0.541	0.821
NA	10	1000	0.052	0.151	0.465	0.761
NA	15	500	0.178	0.441	0.867	0.988
NA	15	1000	0.134	0.370	0.823	0.982
NA	20	500	0.319	0.678	0.977	1.000
NA	20	1000	0.258	0.611	0.966	0.999
	Alignment length = $500nt$					
0.01	10	500	0.145	0.282	0.634	0.885
0.01	10	1000	0.103	0.218	0.560	0.837
0.01	15	500	0.356	0.601	0.922	0.991
0.01	15	1000	0.287	0.525	0.902	0.994
0.01	20	500	0.606	0.846	0.992	1.000
0.01	20	1000	0.537	0.801	0.988	1.000
0.002	10	500	0.261	0.521	0.874	0.985
0.002	10	1000	0.192	0.437	0.828	0.976
0.002	15	500	0.566	0.860	0.997	1.000
0.002	15	1000	0.483	0.812	0.993	1.000
0.002	20	500	0.835	0.980	1.000	1.000
0.002	20	1000	0.929	0.971	1.000	1.000
	Alignment length = $1000nt$					
0.01	10	500	0.114	0.252	0.601	0.862
0.01	10	1000	0.079	0.193	0.526	0.811
0.01	15	500	0.252	0.518	0.867	0.995
0.01	15	1000	0.195	0.442	0.868	0.990
0.01	20	500	0.454	0.760	0.985	1.000
0.01	20	1000	0.379	0.702	0.978	1.000
0.002	10	500	0.177	0.391	0.783	0.958
0.002	10	1000	0.130	0.319	0.726	0.937
0.002	15	500	0.366	0.706	0.978	1.000
0.002	15	1000	0.297	0.642	0.967	1.000
0.002	20	500	0.599	0.899	0.999	1.000
0.002	20	1000	0.528	0.863	0.997	1.000

Table 3.2: The average proportion of distinct gene trees (number of distinct gene trees divided by number of loci) given species trees of 15 and 20 taxa for the sample sizes of 500 and 1000 loci. The entry NA in the coluumn for θ means that known gene trees were used rather then estimated from DNA sequences. The extinction parameter $\mu > 0$ throughout.

ciation events occurring near the present compared to a process with no extinction. The numbers of distinct gene trees and the probabilities of uniqueness were lower when there was extinction, especially for lower speciation rates (Tables 3.3, 3.4, 3.5, 3.6). This is consistent with the observation in Degnan and Salter (2005) that short branches near the root of the species tree tend to lead to higher probabilities of gene trees that do not match the species tree topology. Similarly, pairwise RF and RF-ST distances decreased when extinction was added. Thus, extinction as a biological event effects gene tree heterogeneity.

μ	λ	θ	num. uniq	sd. uniq	RF-ST	pairwise RF	p
0	0.10	NA	217.05	156.10	0.12	0.17	0.00
0	0.20	NA	565.71	222.24	0.22	0.32	0.00
0	0.50	NA	959.64	58.19	0.44	0.58	0.10
0	1.00	NA	999.10	2.61	0.62	0.77	0.72
0	0.10	0.01	308.17	194.72	0.15	0.21	0.00
0	0.20	0.01	659.86	215.59	0.26	0.36	0.00
0	0.50	0.01	978.11	39.33	0.47	0.62	0.22
0	1.00	0.01	999.65	1.11	0.66	0.80	0.85
0	0.10	0.002	493.83	236.58	0.21	0.27	0.00
0	0.20	0.002	864.59	145.28	0.35	0.47	0.02
0	0.50	0.002	998.48	4.28	0.59	0.75	0.65
0	1.00	0.002	999.98	0.14	0.76	0.89	0.98
$.5\lambda$	0.10	NA	134.41	136.47	0.09	0.14	0.00
$.5\lambda$	0.20	NA	370.50	239.13	0.17	0.24	0.00
$.5\lambda$	0.50	NA	823.06	194.92	0.34	0.46	0.02
$.5\lambda$	1.00	NA	981.64	38.77	0.52	0.66	0.37
$.5\lambda$	0.10	0.01	195.07	184.97	0.12	0.17	0.00
$.5\lambda$	0.20	0.01	442.09	253.85	0.19	0.27	0.00
$.5\lambda$	0.50	0.01	868.12	163.89	0.37	0.49	0.07
$.5\lambda$	1.00	0.01	990.05	23.01	0.55	0.69	0.45
$.5\lambda$	0.10	0.002	297.29	232.87	0.16	0.21	0.00
$.5\lambda$	0.20	0.002	642.03	264.16	0.26	0.35	0.00
$.5\lambda$	0.50	0.002	967.30	65.66	0.48	0.62	0.35
$.5\lambda$	1.00	0.002	999.65	1.00	0.67	0.81	0.84

Table 3.3: Summary statistics for species of 15 taxa for sample size of 1000 and DNA sequence length of 1000. NA means that known gene trees were used instead of being estimated from sequences. Here, num. uniq is the average number of unique gene trees, sd. uniq is the standard deviation of the number of unique gene trees, RF-ST is the average RF-ST distance, pairwise RF is the average pairwise RF distance, and p is the probability that all gene trees have a distinct topology in the sample.

μ	λ	θ	num. uniq	sd. uniq	RF-ST	pairwise RF	p
0	0.10	NA	138.970	88.883	0.117	0.174	0.000
0	0.20	NA	320.710	106.361	0.221	0.318	0.000
0	0.50	NA	487.430	20.831	0.435	0.583	0.280
0	1.00	NA	499.780	0.629	0.622	0.768	0.860
0	0.10	0.01	189.590	105.828	0.151	0.213	0.000
0	0.20	0.01	98.419	0.254	0.356	0.010	
0	0.50	0.01	493.220	12.800	0.470	0.617	0.370
0	1.00	0.01	499.960	0.197	0.657	0.797	0.960
0	0.10	0.002	286.080	117.982	0.204	0.285	0.000
0	0.20	0.002	451.470	58.792	0.343	0.467	0.090
0	0.50	0.002	499.640	1.150	0.593	0.746	0.810
0	1.00	0.002	500.000	0.000	0.765	0.887	1.000
$.5\lambda$	0.10	NA	89.250	80.741	0.092	0.135	0.000
$.5\lambda$	0.20	NA	220.540	124.154	0.168	0.240	0.000
$.5\lambda$	0.50	NA	433.360	83.443	0.339	0.464	0.050
$.5\lambda$	1.00	NA	494.110	14.083	0.516	0.664	0.580
$.5\lambda$	0.10	0.01	126.220	102.790	0.122	0.169	0.000
$.5\lambda$	0.20	0.01	258.990	127.993	0.194	0.269	0.000
$.5\lambda$	0.50	0.01	433.570	78.689	0.366	0.492	0.060
$.5\lambda$	1.00	0.01	497.320	6.891	0.546	0.693	0.650
$.5\lambda$	0.10	0.002	182.860	121.519	0.155	0.213	0.000
$.5\lambda$	0.20	0.002	353.210	122.432	0.256	0.351	0.040
$.5\lambda$	0.50	0.002	489.080	24.805	0.475	0.620	0.460
$.5\lambda$	1.00	0.002	499.870	0.418	0.665	0.808	0.900

Table 3.4: Summary statistics for species of 15 taxa for sample size of 500 and DNA sequence length of 1000. NA means that known gene trees were used instead of being estimated from sequences. Here, num. uniq is the average number of unique gene trees, sd. uniq is the standard deviation of the number of unique gene trees, RF-ST is the average RF-ST distance, pairwise RF is the average pairwise RF distance, and p is the probability that all gene trees have a distinct topology in the sample.

μ	θ	λ	num. uniq	sd. uniq	RF-ST	pairwise RF	p
0	NA	0.100	242.890	111.450	0.123	0.181	0.000
0	NA	0.200	431.790	72.670	0.227	0.324	0.020
0	NA	0.500	499.510	1.527	0.441	0.587	0.830
0	NA	1.000	500	0.000	0.630	0.774	1.000
0	0.010	0.100	311.390	113.228	0.158	0.219	0.000
0	0.010	0.200	381.300	60.349	0.257	0.355	0.000
0	0.010	0.500	499.700	0.927	0.471	0.615	0.850
0	0.010	1.000	500	0.000	0.660	0.798	1.000
0	0.002	0.100	400.390	91.178	0.203	0.282	0.010
0	0.002	0.200	492.820	14.905	0.339	0.459	0.410
0	0.002	0.500	499.490	3.350	0.593	0.744	0.860
0	0.002	1.000	500	0.000	0.775	0.895	1.000
0.5λ	NA	0.100	159.330	111.926	0.092	0.135	0.000
0.5λ	NA	0.200	338.970	118.725	0.171	0.245	0.000
0.5λ	NA	0.500	488.660	27.790	0.347	0.474	0.440
0.5λ	NA	1.000	499.810	1.012	0.526	0.673	0.930
0.5λ	0.010	0.100	226.760	126.166	0.124	0.173	0.000
0.5λ	0.010	0.200	379.940	111.405	0.198	0.275	0.020
0.5λ	0.010	0.500	492.380	21.194	0.373	0.501	0.500
0.5λ	0.010	1.000	499.920	0.367	0.554	0.698	0.940
0.5λ	0.002	0.100	299.530	133.290	0.155	0.216	0.010
0.5λ	0.002	0.200	449.430	78.090	0.259	0.355	0.120
0.5λ	0.002	0.500	499.340	3.520	0.479	0.623	0.890
0.5λ	0.002	1.000	500	0.000	0.673	0.813	1.000

Table 3.5: Summary statistics for species of 20 taxa for sample size of 500 and DNA sequence length of 1000. NA means that known gene trees were used instead of being estimated from sequences. Here, num. uniq is the average number of unique gene trees, sd. uniq is the standard deviation of the number of unique gene trees, RF-ST is the average RF-ST distance, pairwise RF is the average pairwise RF distance, and p is the probability that all gene trees have a distinct topology in the sample.

μ	θ	λ	num. uniq	sd. uniq	RF-ST	pairwise RF	p
0	NA	0.100	411.280	214.176	0.123	0.181	0.000
0	NA	0.200	817.060	169.939	0.226	0.323	0.000
0	NA	0.500	997.660	6.089	0.441	0.587	0.700
0	NA	1.000	999.990	0.100	0.630	0.775	0.990
0	0.010	0.100	515.950	227.150	0.158	0.220	0.000
0	0.010	0.200	882.850	130.451	0.257	0.356	0.010
0	0.010	0.500	956.210	152.598	0.471	0.616	0.640
0	0.010	1.000	1000.000	0.000	0.660	0.799	1.000
0	0.002	0.100	741.240	204.552	0.204	0.282	0.000
0	0.002	0.200	976.240	44.367	0.339	0.460	0.270
0	0.002	0.500	999.950	0.219	0.593	0.745	0.950
0	0.002	1.000	999.980	0.141	0.593	0.746	0.980
0.5λ	NA	0.100	258.100	205.427	0.093	0.135	0.000
0.5λ	NA	0.200	611.000	251.043	0.171	0.246	0.000
0.5λ	NA	0.500	965.540	73.200	0.347	0.474	0.250
0.5λ	NA	1.000	999.460	2.645	0.526	0.674	0.880
0.5λ	0.010	0.100	378.680	240.341	0.125	0.173	0.000
0.5λ	0.010	0.200	702.280	240.798	0.198	0.276	0.010
0.5λ	0.010	0.500	978.150	54.421	0.373	0.501	0.380
0.5λ	0.010	1.00	999.750	1.019	0.554	0.699	0.890
0.5λ	0.002	0.100	527.790	269.594	0.155	0.216	0.000
0.5λ	0.002	0.200	863.430	182.156	0.259	0.356	0.080
0.5λ	0.002	0.500	997.480	12.194	0.479	0.623	0.760
0.5λ	0.002	1.000	1000.000	0.000	0.673	0.815	1.000

Table 3.6: Summary statistics for species of 20 taxa for sample size of 1000 and DNA sequence length of 1000. NA means that known gene trees were used instead of being estimated from sequences. Here, num. uniq is the average number of unique gene trees, sd. uniq is the standard deviation of the number of unique gene trees, RF-ST is the average RF-ST distance, pairwise RF is the average pairwise RF distance, and p is the probability that all gene trees have a distinct topology in the sample.

3.2.4 *Gene tree estimation error*

Gene tree estimation error is often reported as an important factor that makes species tree inference more challenging (Huang et al., 2010; DeGiorgio and Degnan, 2014; Roch and Warnow, 2015; Xi et al., 2015; Roch et al., 2019; Cai et al., 2021). These simulations also illustrate that gene tree estimation error increases gene tree heterogeneity. The effect of gene tree estimation error can be seen by comparing measures of heterogeneity for known versus estimated trees, and lower versus higher quality estimated trees due to shorter alignments and lower values of θ (which causes lower information in the DNA alignments).

In Figure 3.2, gene tree estimation error is introduced in the second column, and increased in the third column by shortening the alignment length, resulting in higher probabilities that all gene trees are unique. For example, with 15 taxa, 500 loci per sample, and $\lambda = 0.5$ with no extinction, having mutation change from none to $\theta = 0.01$ to $\theta = 0.002$ resulted in the proportion of simulated data sets with all gene tree topologies being unique change from 0.28 to 0.81 to 0.97. The sensitivity of the probability that all topologies were unique varied considerably depending on both λ and θ, as well as the number of loci. A comparison of Figures 3.1 and 3.4, and also Figures 3.2 and 3.3, show that increasing θ from 0.002 to 0.01 (which decreases gene tree estimation error), decreases the number of unique gene tree topologies as well as the probability that all gene tree topologies are unique.

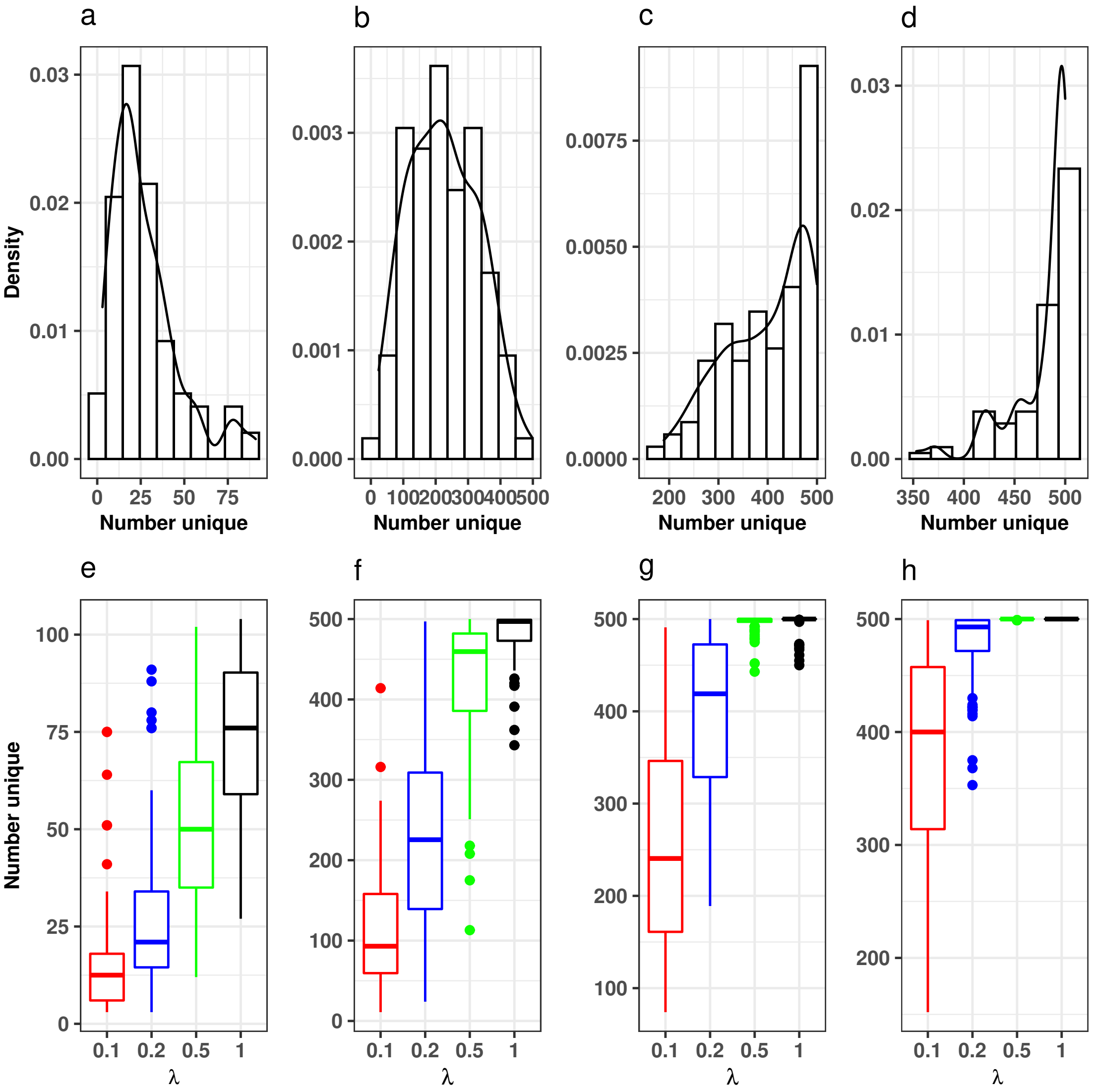

Figure 3.4: Plots (a)–(d) are distributional plots for the number of unique gene tree topologies when there are 5, 10, 15 and 20 species, respectively, for a sample size of 500 gene trees with $\mu = 0$, $\lambda = 0.2$, and $\theta = 0.010$. Plots (e)–(h) are boxplots for the number of unique gene trees as a function of the speciation rate λ when $\mu = 0$ in samples of size 500 gene trees for $n = 5, 10, 15,$ and 20 species, respectively.

3.2.5 *Correlation*

We investigated the relationships of the pairwise RF distance (the average RF distance between all pairs of gene trees) among the gene trees and the RF distance between the gene trees and species tree (average RF distance between the gene tree and species tree) through a combination of simulations and biological data analyses. Across all the numbers of taxa and λ's in the simulation, RF and RF-ST have a strong positive correlation (Figure 3.5 and Table 3.7). Also, both pairwise RF and RF-ST increase with λ and rapidly when $\mu = 0$ under all the simulation conditions (Table 3.3).

There are many methods for inferring species trees from collections of gene trees (Liu et al., 2009; Xu and Yang, 2016). One of the measures of the levels of discordance of gene trees and species tree is the RF distance between the gene trees and the species tree. A high value of this measure indicates a high probability of discordance. However, reporting the RF distance between the reconstructed gene trees and the species tree requires knowing the actual species tree. While this is possible in simulation studies, it can only be estimated in empirical studies. Although the RF-ST is sometimes used to characterize the level of ILS (e.g. Mirarab et al., 2014), our simulation results indicate that one can use the pairwise RF distance instead of the RF-ST distance since both are strongly and positively correlated (Figure 3.3

and Table 3.5). Also, the simulation results confirmed that both pairwise RF and RF-ST increase with the number of unique trees. Thus, the values of either pairwise RF or RF-ST distances could be used to describe the degree of heterogeneity of the gene trees given a species tree.

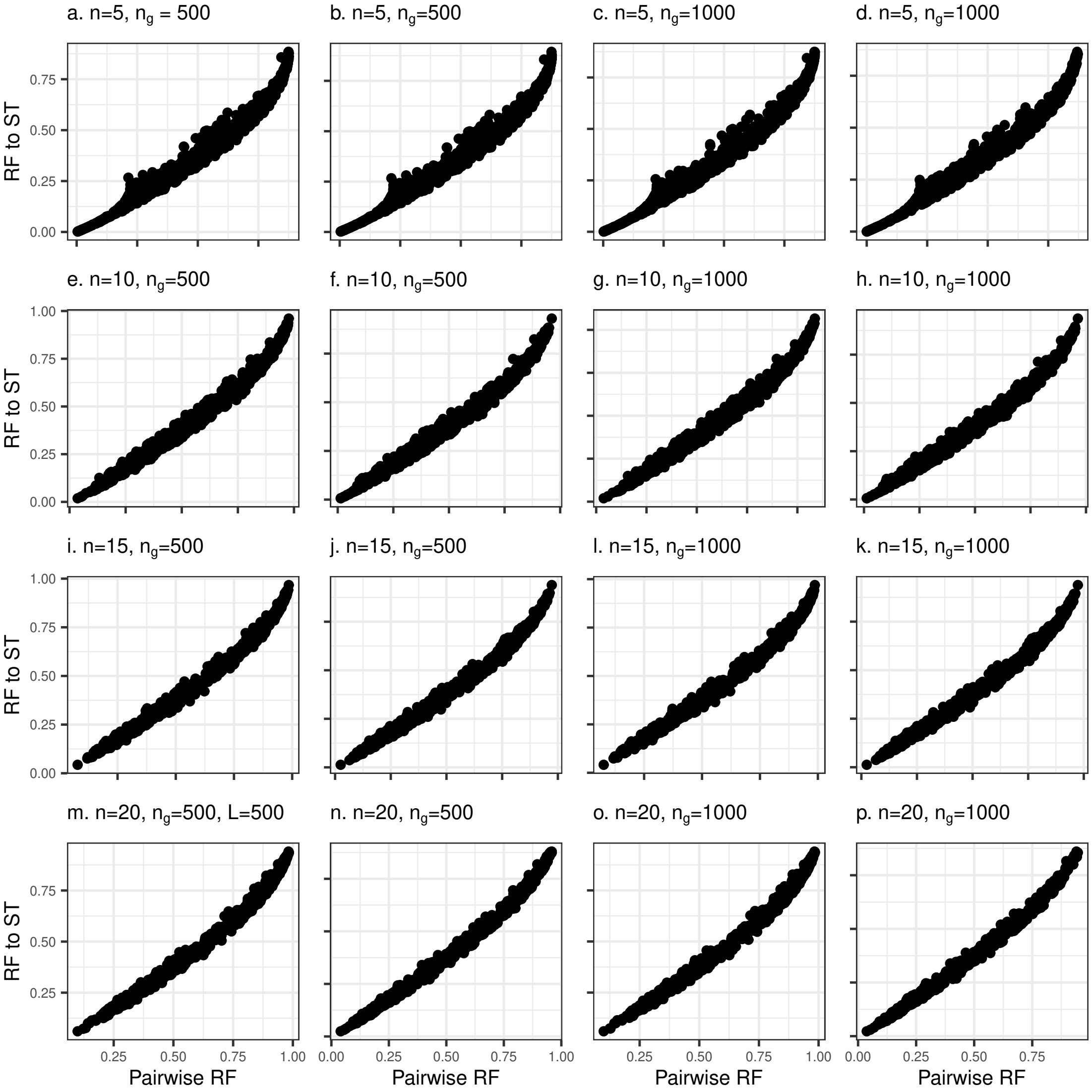

Figure 3.5: The plots of RF distances between estimated gene trees to species trees (RF to ST) versus pairwise RF. Number of taxa increases by row, and columns alternate between alignment lengths of 500 and 1000. For all plots, $\theta = 0.002$.

	$\lambda = 0.1$			
n	μ	θ	$r_{(pRF,RF-ST)}$	$r_{(pRF,nUniq)}$
5	0.000	0.002	0.978	0.887
5	0.050	0.002	0.980	0.841
5	0.000	0.010	0.970	0.838
5	0.050	0.010	0.980	0.803
10	0.000	0.002	0.987	0.980
10	0.050	0.002	0.980	0.956
10	0.000	0.010	0.979	0.959
10	0.050	0.010	0.979	0.921
15	0.000	0.002	0.982	0.924
15	0.050	0.002	0.982	0.971
15	0.000	0.010	0.975	0.986
15	0.050	0.010	0.979	0.980
20	0.000	0.002	0.977	0.772
20	0.050	0.002	0.987	0.864
20	0.000	0.010	0.969	0.948
20	0.050	0.010	0.985	0.975
	$\lambda = 0.5$			
5	0.00	0.002	0.969	0.980
5	0.5λ	0.002	0.965	0.971
5	0.00	0.010	0.976	0.940
5	0.5λ	0.010	0.975	0.931
10	0.00	0.002	0.979	0.818
10	0.5λ	0.002	0.981	0.909
10	0.00	0.010	0.984	0.952
10	0.5λ	0.010	0.982	0.983
15	0.00	0.002	0.979	0.379
15	0.5λ	0.002	0.981	0.576
15	0.00	0.010	0.985	0.630
15	0.5λ	0.010	0.982	0.767
20	0.00	0.002	0.964	NA
20	0.5λ	0.002	0.983	0.233
20	0.00	0.010	0.981	0.497
20	0.5λ	0.010	0.986	0.543

Table 3.7: Correlation coefficients of Pairwise RF and RF-ST, and number of unique gene trees and pairwise RF for $\lambda = 0.1$ and $\lambda = 0.5$, sample size of 500 and DNA sequence length of 500.

3.3 Discussion

This work was motivated by the observation that multilocus studies often had high levels of gene tree heterogeneity, and in particular that every gene tree having a unique topology could seem surprising. For the example given in Salichos and Rokas with 23 taxa and 1070 genes, having every gene tree topology be unique would not be particularly surprising for higher speciation rates, although would be surprising for very low speciation rates. For example, with 20 taxa and 1000 loci, for $\lambda \geq .5$ and either value of μ (0 or $.5\lambda$), the probability of all gene tree topologies being unique was at least 70% with known gene trees, and higher with estimated gene trees, so that it would not be surprising to not have all unique topologies in this setting. For a lower speciation rate, say $\lambda = 0.2$, it would not be be particularly surprising to either have or not have all unique gene tree topologies.

Overall, we see that the curves describing the probabilities of uniqueness rise steeply with λ, so that seeing many distinct gene trees in a data set is informative regarding the speciation rate–high levels of gene tree heterogeneity, at least in the absence of processes such as hybridization and ancestral population structure, are consistent with high speciation rates and less consistent with lower speciation rates. Quality of the inferred gene trees matters as well, since gene trees more accurately constructed (such as with longer alignments) tend to have less heterogeneity.

Chapter 4

Estimation of Birth-Death Parameters

4.1 Birth-Death Parameters

One of the fundamental features of evolutionary dynamics is its branching structure which depends on the rates of speciation and extinction. When we observe the present day species, we can partially only observe the number of species and some of their characteristics, such as, their DNA sequences, but the evolutionary genealogy that produced the species remains unobserved. An interesting question is "how can we infer the parameters behind the evolved process that led to the observed species?" Knowing the rate(s) of the evolutionary process helps us trace and reconstruct the relationships of the present day species going back in time. Also, the branching times of a phylogenetic tree contains information about the Birth-Death models (Nee et al., 1994). At any time point, the macroevolutionary process involves one of the following: (1) speciation (species giving birth to new species), (2) extinction (species dying off) and (3) neither speciation nor extinction (Feller, 2008). The rates at which these events happen are of interest, and there are several proposed methods of inference of these rates (birth-death parameters).

A popular method for estimating the birth-death parameters is either directly from allele frequencies (Tanaka et al., 2006; Stadler, 2011b) or from the species tree estimated from gene trees that were estimated from DNA sequence. Inference of the birth-death parameters based on phylogenetic information depends on several assumptions: the quality of the phylogenetic data, the accuracy with which branch lengths are calibrated to time and the constancy of the speciation and extinction rates within clades (Turelli et al., 2001; Coyne and Orr, 2004; Ricklefs, 2007). However, despite probabilistic theorems on estimating evolutionary parameters, performing statistical inference using genomic data can be very complicated and challenging. Some of reasons for this are: (1) even simple models of evolutionary processes are often mathematically intractable, (2) the branching pattern of the phylogeny depends on many implicitly unobserved factors such as, hybridization and introgression or horizontal gene transfer. In this chapter, we propose to use approximate Bayesian computational (ABC) methods to estimate the parameters from the gene trees.

In the remaining part of this chapter, we describe the birth-death process and likelihood-based method, and a simulation study for ABC methods. We conclude the chapter with a brief discussion of the results from the simulated data.

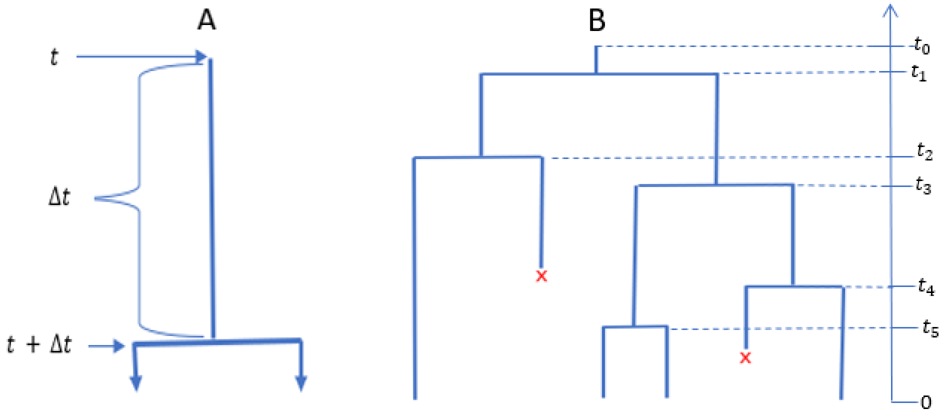

Figure 4.1: Illustration of a model (birth-death process) that shows the lineage that gave rise to the present day species. A). Waiting time to a speciation event; B). A birth-death tree with waiting times, where x denotes extinct species. The values t_1, t_2, t_3, t_4, t_5 are are speciation times.

4.1.1 Birth-death process

The Birth-death process is a continuous-time Markov chain that models how the number of species changes over time. A lineage speciates into new lineage(s) at a rate λ and goes extinct at a rate μ. We assume that $\lambda > \mu$ and $\lambda > 0$, otherwise the birth-death process is at a critical or subcritical situation. That is, the lineages will die out with probability 1. However, even if $\lambda > \mu$, a lineage can go extinct (Stadler, 2013a; Stadler et al., 2013). Of course, the rate of speciation and extinction at any given time depends on the number of species at that time. A common assumption is that these parameters are constant, and have linear relationships with other evolutionary factors. However, in practice the evolutionary process is more complicated and usually arises from implicit complex mechanisms that underlies the biological structure of the species. These factors hinder the use of the birth-death process for the inference of birth-death parameters due to difficulty in performing the statistical estimation (Holmes and Bruno, 2001). Thus, the application of the birth-death process for the inference of the birth-death parameters has been limited to a continuous process. Consequently, much work focuses on the simple linear birth-death process since it is somewhat analytically tractable. Unfortunately, a simple birth-death process may not capture the complicated lineages' relationships.

We assume that in a speciation-extinction process only two things can happen: births, where the number of lineages (species) increases by one; and deaths, where the number of lineages decreases by one. The simplest branching process for species evolution assumes that one species splits only into two or die during any one event; thus, trees cannot have "**hard polytomies**" (multifurcation). Usually, researchers consider a process where each species has constant rates of either speciating (giving birth) or going extinct (dying). The process starts with a single lineage at some time t_0 in the past and has the probability of either speciating or going extinct. The branching process continues in both new species going forward in time. Typically, there are waiting times before the next event (either speciation or extinction), see Figures 4.1(A) and 4.1(B). A special case of a birth-death process is the Poisson process, in which the number of events can only increase over time.

A pure birth model ($\lambda > 0$ and $\mu = 0$) of the birth-death process is a Poisson process.

Let $N(t)$ be the number of lineages at time t, including those that have gone extinct and not observed in the pruned phylogeny. Also, let L be the number of lineages in a clade. According to Ricklefs (2007), the expected value of L is: $E[L] = e^{(\lambda-\mu)t}$ and

$$N(t) = \frac{\lambda E[L] - \mu}{\lambda - \mu} = \frac{\lambda e^{(\lambda-\mu)t} - \mu}{\lambda - \mu} \tag{4.1}$$

and when the extinct lineages are pruned from a phylogeny, the number of lineages ancestral to present day species is

$$N_A(t) = \frac{\lambda e^{(\lambda-\mu)T} - \mu}{\lambda e^{(\lambda-\mu)(T-t)} - \mu} = \frac{N(T)}{N(T-t)} \tag{4.2}$$

where T is the age of the phylogeny. Also, the difference between $\ln N(t)$ and $\ln N_A(t)$ (Harvey et al., 1994; Ricklefs, 2007) is approximately $-\ln \frac{\lambda-\mu}{\lambda}$.

A birth-death process in macroevolution only considers the total number of the species and do not keep track of the ancestor. The rate of speciation and extinction at any time is a function of the number of the tips of the species. We can understand the behavior of birth-death processes in phylogenetic settings if we consider the waiting time between successive speciation and extinction events in the phylogeny. Suppose we consider a single lineage that exists at time t. The next event is either a speciation event, splitting the lineage into two (Figure 4.1A), or an extinction event marking the end of that lineage (Figure 4.1B). Under a birth-death process, the expected waiting time for an event follows an exponential distribution, with parameters either λ or μ. Thus, the expected waiting time for the next event is exponential with parameter $\lambda + \mu$, and the probability that the next event is speciation is $\frac{\lambda}{\lambda+\mu}$ and the probability that the next event is extinction $\frac{\mu}{\lambda+\mu}$.

When more than one species is 'alive' at the time t, speciation occurs with the rate λ_N and extinction with the rate μ_N. In the classical simple linear birth-death process, $\lambda_N = N\lambda$ and $\mu_N = N\mu$. In a general birth-death process, λ_N and μ_N can be any function of N and the waiting time to the next event. The distribution of the waiting time is exponential with parameter $N(\lambda+\mu)$. The rate parameter of this exponential distribution gets larger and larger as the number of species increases, and the expected waiting times across all species get shorter and shorter as the number of species increases.

One popular approach in the literature, for inference of birth-death parameters using birth-death model is to first estimate the net diversification rate

$$r = \lambda - \mu \tag{4.3}$$

and the relative speciation or extinction rate (Stadler, 2011b)

$$\epsilon = \frac{\lambda}{\mu} \text{ (or } \frac{\mu}{\lambda}) \tag{4.4}$$

The diversification rate can be estimated from the present day species diversity and its age (Wilson, 1983; Magallon and Sanderson, 2001), and if extinction is negligible,

$$\hat{r} = \hat{\lambda} = \frac{log(n)}{t}$$

and

$$\hat{r} = \hat{\lambda} = \frac{log(n) - log2}{t}$$

for a stem group age and for a crown group age, respectively. Here, n is the recent species diversity, t is the age of the tree in millions of year. A crown group is the group that includes all the present day species of a clade plus all the extinct descendants back to the common ancestor of all the present day species. The stem group includes all species that are not part of the crown group. That is, every member of the stem group that has gone extinct (Figure 4.2). Then simultaneously estimate birth-death parameters from equations (4.3) and (4.4).

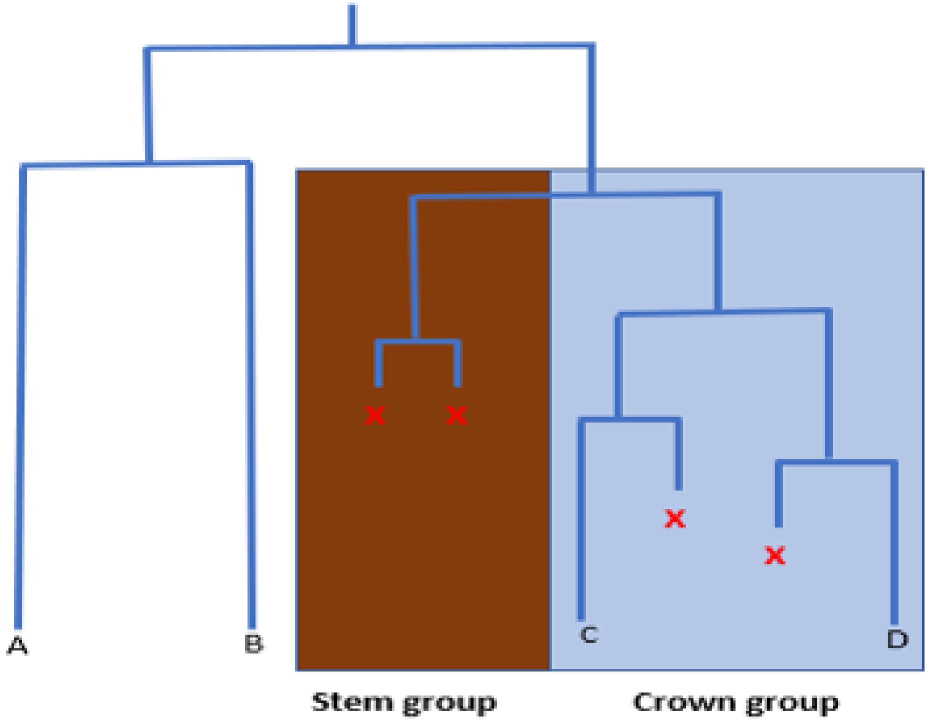

Figure 4.2: Illustration of a tree with Crown and Stem groups.

The birth-death process can be applied in many other settings. For example,it can be use to study infectious disease dynamics in a finite population, where the number of infected individuals is the quantity of interest (Andersson and Britton, 2012). In molecular evolution, the birth-death process can be used to model the insertion and deletion of nucleotides in a DNA sequence. Further, it is commonly used to model quantities of interest in allele frequencies, selection, or coalescence studies (Kingman, 1982; Moran, 1958).

Properties of the Birth-Death Model

Suppose we consider a small interval, Δt, and assume that the length of this interval is so short that it can only contain either one event or no event. Therefore, the probability of an occurrence of an event in this interval with rate r is approximately $r\Delta t$. The probability of speciation event over this interval is:

$$P_{speciation} = \lambda_N \Delta t + o(\Delta t) = N(t)\lambda\Delta t + o(\Delta t)$$

The probability of extinction event in this interval is:

$$P_{extinction} = \mu_N \Delta t + o(\Delta t) = N(t)\mu\Delta t + o(\Delta t)$$

and the probability of more than one event is:

$$P_{more\ than\ one\ event} = o(\Delta t)$$

where $N(t)$ is the number of surviving species at time t. For the term, $o(\Delta t)$, it means that it has smaller order than Δt, so that

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0.$$

This implies that the probability of more than one birth in this interval Δt is negligibly small.

Combining the above equations, we obtain that the probability of neither birth nor death in the interval $(t, t + \Delta t)$ is

$$P_{No\ event} = 1 - (\lambda_N + \mu_N)\Delta t + o(\Delta t) \tag{4.5}$$

Expected Number of Living Species

Suppose the total number of living species at time t is $N(t)$. For a very small interval of time Δt, the expected value of $N(t)$ is

$$E\left[N(t + \Delta t)\right] = N(t) + N(t)\lambda\Delta t - N(t)\mu\Delta t \tag{4.6}$$

Subtracting $N(t)$ from both sides of (4.6), dividing by Δt and then taking the limit as Δt tends to 0, We obtain a differential equation:

$$\frac{dN}{dt} = N(t)(\lambda - \mu)$$

Without loss generality, we assume that there are N_0 species at time $t = 0$. Then, the solution to the differential equation with an initial value condition ($N(0) = N_0$) is

$$N(t) = N_0 e^{(\lambda - \mu)t} \tag{4.7}$$

This is the expected number of species over the time under a birth-death model, and this number has exponential growth or decay if $\lambda > \mu$ or $\lambda < \mu$, respectively. Also, it is constant as $\lambda = \mu$, (Figure 4.3). For a pure birth process, we have

$$N(t) = N_0 e^{\lambda t}$$

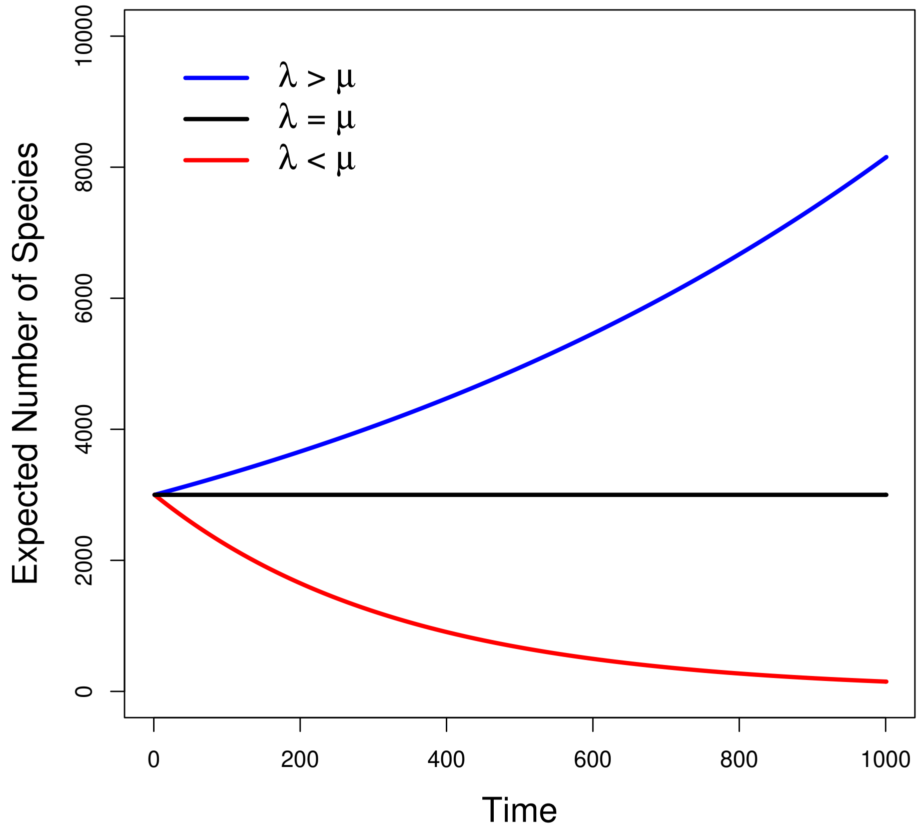

Figure 4.3: Expected number of species under a birth-death model with $N_0 = 3000$.

4.1.2 Maximum Likelihood Method

A likelihood is a probability (or density) of observing the data (phylogeny) given the proposed values of the parameters. The maximum likelihood method for the birth-death parameters is to maximize the likelihood of observing a species tree from the tree's branch lengths. We assume that the tree is ultrametric (i.e., the total root to tip distance is the same for every species). To use the maximum likelihood method to estimate the parameters, one needs to write down the likelihood functions of the waiting times between the speciation and extinction events. Knowing the probability density function allows making inferences about the maximum likelihood of the birth-death parameters for a given tree (reconstructed) by maximizing the density function over the parameters.

The convention is to assume that the tree begins at time t_1, the root node, with a pair of species (Stadler, 2013a) and the initial lineages survive to the present day. The observed phylogenetic tree is a pruned tree (i.e., ignoring extinct species). A tree with n tips has $n-1$ speciation times, denoted as $t_1, t_2, ..., t_{n-1}$ and $t_i > t_{i-1} > 0$ (Figure 4.1), for $i = 1, ..., n$, with the present day time being 0 (Nee et al., 1994; Stadler, 2010, 2013a,b). The times of bifurcations of phylogeny provide some information about the speciation and extinction rates, and these rates can be estimated from the reconstructed species tree (Stadler and Steel, 2012a). The speciation time is measured backward in time from the present day. Note, we can use both complete and incomplete phylogenies to estimate speciation and extinction rates (Stadler and Steel, 2012a; Rabosky et al., 2007), especially, when we have complete sampling; that is, all the n tips are represented in the tree.

The general idea is to assign probabilities to the tips of the tree and update them going backward in time to the root (Nee et al., 1994; Maddison et al., 2007; Stadler, 2013a). Then getting at the root, we obtain the probability of observing the tree given the model, and this probability at the root is the likelihood. To do this, we need to keep track of two things: (1) the probability $S_N(t)$ that a lineage at time t in the past speciated and gave birth to node N that survives to the present day species and (2) the probability $E(t)$ that a lineage evolves at some point but went extinct before the present day; the lineage starting at time t, leaves no descendants at the present day. Going backward in time from the tip to the root, $S(t_{root})$ is the full likelihood for a single lineage. Note that $S_N(t)$ and $E(t)$ depend on each other because the probability of observing a lineage on a tree depends on the extinction probability of that lineage and vice visa, and these are calculated backward in time. We note that $S_N(t_n) = 1$ and $E(t_n) = 0$ at the tip because observing a tip at present day means that it does survive and the probability is 1.

Given the speciation rate λ, the time it takes a lineage to bifurcate has an exponential distribution with a mean of $\frac{1}{\lambda}$ (Mooers et al., 2012; Mooers and Heard, 1997; Nee et al., 1994; Stadler, 2013a). Suppose we consider a branch of a tree with no nodes, going back in time, we know that (1) the lineage did survive, and (2) if speciation occurred, the lineage(s) that bifurcated did go extinct and did not survive to the present day. These scenarios are represented in the following equations (Maddison et al., 2007)

$$\frac{dS(t_n)}{dt} = -(\lambda + \mu)S(t_n) + 2\lambda E(t)S(t_n) \tag{4.8}$$

The first term of the equation, $-(\lambda + \mu)S(t_n)$, is the rate of no speciation or extinction, while the second term $2\lambda E(t)S(t_n)$ is the rate of speciation followed by the extinction of either daughter lineage.

$$\frac{dE(t)}{dt} = \mu - (\lambda + \mu)E(t) + \lambda E(t)^2 \tag{4.9}$$

The three terms in the equation are the three possible ways a lineage might not survive to the present day. That is, either goes extinct in the interval considered or survives the interval but goes extinct sometime later, or it splits in the interval, but both descendants do not survive to the present day. Note, $E(t)$ only depends on the time and not on the number of the tips of the tree, and each descendant that goes extinct is independent, hence the term $E(t)^2$.

Solving the ordinary differential equations with initial conditions: $E(0) = 0$ and $S_N(0) = 1$ (Maddison et al., 2007), then

$$E(t) = 1 - \frac{\lambda - \mu}{\lambda - (\lambda - \mu)e^{(\lambda - \mu)t}} \tag{4.10}$$

and

$$S_N(t) = e^{-(\lambda-\mu)(t-t_N)} \frac{(\lambda - (\lambda - \mu)e^{(\lambda-\mu)t_N})^2}{(\lambda - (\lambda - \mu)e^{(\lambda-\mu)t})^2} S_N(t_N) \tag{4.11}$$

where t_N is the time depth of node N. Considering each branch's likelihood and taking product over all $2n - 2$ branches, the likelihood of observing the tree given λ and μ is proportional to (Maddison et al., 2007):

$$L(t_1, ..., t_n|\lambda, \mu) = \lambda^{n-1} \cdot \prod_{j=1}^{2n-2} e^{(\lambda-\mu)(t_{j,t}-t_{j,b})} \cdot \frac{(\lambda - \mu e^{-(\lambda-\mu)t_{j,t}})^2}{(\lambda - \mu e^{-(\lambda-\mu)t_{j,b}})^2} \tag{4.12}$$

where $t_{j,b}$ is the time at the base of the j^{th} branch and $t_{j,t}$ is the time at the node on the j^{th} branch length nearest to the observed species (present day taxon).

This procedure only entails knowing the likelihood equation of observing the set of the speciation times $t_1, t_2, ..., t_{n-1}$, the number of the tips n and then obtaining the MLE of the parameters from the equation. However, as the number of species increases, the likelihood equation become computational expensive and sometimes intractable. Estimating the parameters from gene trees entails combining all the likelihoods of the gene trees used in in the inference of the species tree together.

For a pure birth model ($\mu = 0$), we can write the likelihood function as

$$L(t_1, ..., t_n|\lambda) = \lambda^{n-1} \cdot \prod_{j=1}^{n-1} e^{-\lambda t_j} \tag{4.13}$$

and the maximum likelihood estimate of λ is:

$$\hat{\lambda} = \frac{n-1}{\sum_{j=1}^{n-1} t_j} \tag{4.14}$$

where $\sum_{j=1}^{n-1} t_j$ is the sum of the branch lengths. However, the resulting likelihood equation (4.12) may be complex, complicated or even intractable. Thus, in this project, we propose to use the ABC algorithm to infer the birth-death parameters assuming that the trees are ultrametric. In particular, we compare the results of the ABC method and maximum likelihood method for the pure birth model.

4.2 Approximate Bayesian Computation (ABC) Methods

Approximate Bayesian computation is a class of computational methods in statistics rooted in the Bayesian method and is used to estimate the posterior (or predictive) distributions of model parameters. ABC algorithms are used to approximate the likelihood function by simulations, and the outcomes are compared with the observed data. This method is popularly used for inference when the likelihood function can be simulated but is not analytically tractable. ABC methods are popular in the biomedical sciences, particularly in genetic and phylogenetic studies (Tanaka et al., 2006; Csilléry et al., 2010; Fan and Kubatko, 2011; Stadler, 2011b; Kutsukake and Innan, 2014; Veeramah et al., 2015; Janzen et al., 2015; Alanzi and Degnan, 2017).

ABC methods provide a means of performing inferences when confronted with unreasonably complex and complicated models and the methods rely on summary statistics. A comparison of how similar the simulated data set $\boldsymbol{Y}$ to the observed data set $\boldsymbol{X}$ is undertaken by computing a distance $\boldsymbol{d(X,Y)}$. Parameter samples that produce simulated data sets that are 'very close' to the observed data set X are collected and kept as samples from the posterior distribution. That is, a parameter that produces simulated data that has its distance from the observed data less than the **tolerance** (ϵ) is accepted as part of the posterior.

4.2.1 ABC Rejection Algorithm

The most basic set in ABC is a set of parameter points that is first sampled from the prior distribution. Then the sampled parameter points λ, are used to simulate data Y under the model specified by λ. The sampled parameter value is rejected if the simulated Y is too different from the observed data X. That is, Y is accepted with tolerance ϵ if:

$$d(X, Y) \leq \epsilon$$

This value (ϵ) can be strict or not, depending on the problem, and also can be chosen based on prior experience. Sometimes, the optimal tolerance value is up for debate. Practically, the probability of generating a data set Y with small (or very small) distance to data set X decreases as the sample size increases, and this might lead to a decrease in the computational efficiency of the rejection algorithm (Sunnåker et al., 2013). The popular approach is to replace Y and X with sets of summary statistics (summaries of the data that do not lose any information useful for doing inference), $S(Y)$ and $S(X)$, respectively, (Casella and Berger, 2021). Typically, the summary statistics have smaller dimensions and contain information in X and Y, respectively. Further, if the the relevant information in X and Y are well captured by $S(X)$ and $S(Y)$, this dimensional reduction does not introduce any error or bias (Didelot et al., 2011; Sunnåker et al., 2013). The distance measure $d(X, Y)$ is also replaced with $d(S(X) - S(Y))$, which determines the level of discrepancy between the data set $S(X)$ and data set (Y) based on a chosen metric. Common and popular metrics are:

- **the sum of absolute difference**

$$d(S(X), S(Y)) = \sum_{k \in K} |S(X)_k - S(Y)_k|$$

- **the sum of squared difference**

$$d(S(X), S(Y)) = \sum_{k \in K} [S(X)_k - S(Y)_k]^2$$

- **the Kullback-Leibler divergence (KL)**

$$d(S(X), S(Y)) = \sum_{k \in K} S(X)_k \log\left(\frac{S(X)_k}{S(Y)_k}\right)$$

Then the acceptance criterion is:

$$d(S(X), S(Y)) \leq \epsilon.$$

The result of the ABC rejection algorithm is a sample of parameter values distributed according to the posterior distribution and obtained without evaluating the likelihood function explicitly. In this project, the metric we used is the pairwise Robinson-Flouds (RF) distance (Robinson and Foulds, 1981). RF distances of sampled gene trees from species tree simulated from the prior that corresponded to small distances compared to the observed RF are used as a basis for an estimated posterior distribution for the parameter (i.e., the speciation rate). Basically, in Bayesian statistics, the aim is to determine the posterior distribution of the parameter given the data. ABC approach estimates the posterior distribution given that the summary statistic is close to the data, and this approximates the posterior well when the summary statistic contains the relevant information in the data. So, data are simulated under a range of values of the parameter λ. At each step, if the data that is produced matches or is very close to the observed data, observed RF, the parameter value that is being used to generate the data is 'accepted.' The set of accepted parameter values is then used to approximate the posterior distribution. That is, for each simulated RF' that is identical or very to the observed data observed RF, the generating parameter λ values are stored (that realization is ***accepted***) and used to construct a posterior distribution for the parameters. In this project, we chose the best 500 simulated RF that were very closest to the observed RF.

The main advantage of this method is that, for the most complicated phylogenetic settings, it is far easier to simulate than to calculate. In fact, many models of evolution lead to distributions for which direct calculation is complicated or impossible, but which, given improvements in computational efficiency, can be relatively easily simulated. So, this leads to the somewhat easy development of rejection algorithms, with evolutionary models, for the purposes of inference.

In fact, we spend a lot of time generating parameter values from the prior distribution, only to discover that they rarely lead to data that is 'very close' the observed RF. The 'acceptance rate' of such algorithms is so low that it takes an unreasonable amount of time to collect a large set of accepted parameter values. Thus, we use an alternative method, the best βN of the parameters that produced data are very close to the observed data in the absolute distance. We note that since the summary statistics we used may not be sufficient statistics, the resulting predictive distribution is an approximation of the true posterior, and the closeness of the approximation is, a priori, unknown.

Further, we hope that the ABC method will be useful in this project since the likelihood function of gene trees from a species tree is complicated and likely to be intractable. The details of our ABC algorithm are given in the method section below. We simulated the gene trees from 10-taxon and 15-taxon species trees.

4.2.2 Methods

We first simulate from the prior distribution for the parameter (in this case, a speciation rate, λ), then simulate data from the parameter (a species trees and gene trees), using TreeSim (Stadler, 2019) and hybrid-Lambda (Zhu et al., 2015), compute a pairwise RF distance of the simulated and observed gene trees, and then record an absolute difference between the RF distances. In this project, the observed and simulated data consist of species trees, gene trees from the species tree, and pairwise RF distances among the gene trees. We apply the method to 10-taxon and 15-taxon trees.

For 10-taxa, there are 34,459,425 possible rooted tree topologies. Based on the observed parameter, λ, we randomly generate a species tree and $n_g = 500$ gene trees from this species tree and then compute the average pairwise RF distance among the gene trees.

Also, for the simulated data, we first generate a vector of length 10,000 from the prior distribution as speciation rates and simulate species trees with each data point from the vector. Further, samples of gene trees with 500 loci per species tree are generated and the sum of the RF distances is computed. The details of the algorithm are given below.

Algorithm

1. Simulate a species trees with observed speciation rate (λ) and extinction rate ($\mu = 0$) using TreeSim (Stadler, 2019), and then generate gene trees from the species trees using hybrid-Lambda (Zhu et al., 2015) and compute the sum of pairwise RF distance among the gene trees; call this *observed-RF*.

2. start with $i = 1$.

3. Simulate a species tree from the prior distribution of λ using TreeSim.

4. sample gene trees from the species tree generated with the prior using hybrid-Lambda; then calculate the pairwise RF distance of the gene trees and call it *simulated-RF*$_i$.

5. Calculate $D_i = |\textit{simulated-RF}_i - \textit{observed-RF}|$

6. Increment i by 1 and repeat steps (2)-(5) N times

7. Take the smallest βN values from step (5), then retain the speciation rates corresponding to these smallest distances. These rates estimate the posterior (or predic-

tive) distribution.

Here, β is chosen to be a small number so that only those simulated speciation rates with simulated RF distance close to the observed RF distance are retained. We used $\beta = 0.05$ and $\beta N = 500$ so that the posterior is formed from the 500 best speciation rates. We follow Alanzi and Degnan (2017); Fan and Kubatko (2011) in accepting a fixed number of speciation rates instead of the smallest distances. This approach is common in practice in ABC, and it is equivalent to using a threshold that is based on a quantile of the simulated distribution of distance (Beaumont et al., 2002). The approach we used in this project, that is, the best $100 \times \beta\%$ of the speciation rates, leads to a fixed number N of data sets to simulate, and it is easier to use in practice since the length of the simulation is known in advance, making easier to plan the length of time needed for computation. In contrast, the fixed threshold approach leads to a random number of iterations needed to obtain a fixed number of accepted parameters to estimate the posterior distribution. We estimate the birth parameters by summarizing the posterior distribution of the speciation rates. To the perform all the computations, we employs several scripts, including R codes; details are given in the Appendix.

4.2.3 Simulation and Sensitivity Analysis

For 10 taxon and 15 taxon trees, we also used the Algorithm above to simulate one species tree using TreeSim (Stadler, 2019) under a pure birth model ($\mu = 0$) with a speciation parameter from a prior, exponential distribution with a mean of 0.5. The choice of the prior was based on a randomly selected tree pendant edge that has an exponentially distributed length with parameter 2λ (Stadler and Steel, 2012a). Also, we simulated the specie trees with birth parameters from an exponential distribution with mean 1 and a beta distributions with parameters, $\beta = 2$ and $\alpha = 2$, to investigate how sensitive the predictive distribution is to the shape and center of the prior distributions. The predictive distribution estimated using ABC is summarized using speciation rates that their species trees generate a sample of gene trees of the best 500 RF distances. The best 500 speciation rates that produced the species trees that the sum of the pairwise RF among the gene trees sampled from them are close to the observed RF. Note that the speciation rate for the observed species tree that its gene trees produced the observed RF distance is 0.5 Using the RF distance to summarize a distribution of trees is fairly common in phylogenetics (Mirarab et al., 2014). Further, we outlined different proposed steps for investigating the performance of the ABC method other methods on Figure 4.4 Some of the steps are:

- maximum likelihood (ML) using STEM (Kubatko et al., 2009) to estimate the speciation parameter directly from branch lengths of the species tree.

- maximum likelihood on the branch lengths of an estimated species tree from gene trees generated with hybrid-Lambda (Zhu et al., 2015).

- ABC to estimate the speciation rate from the gene trees, and so on.

Most of the steps are planned for the future studies.

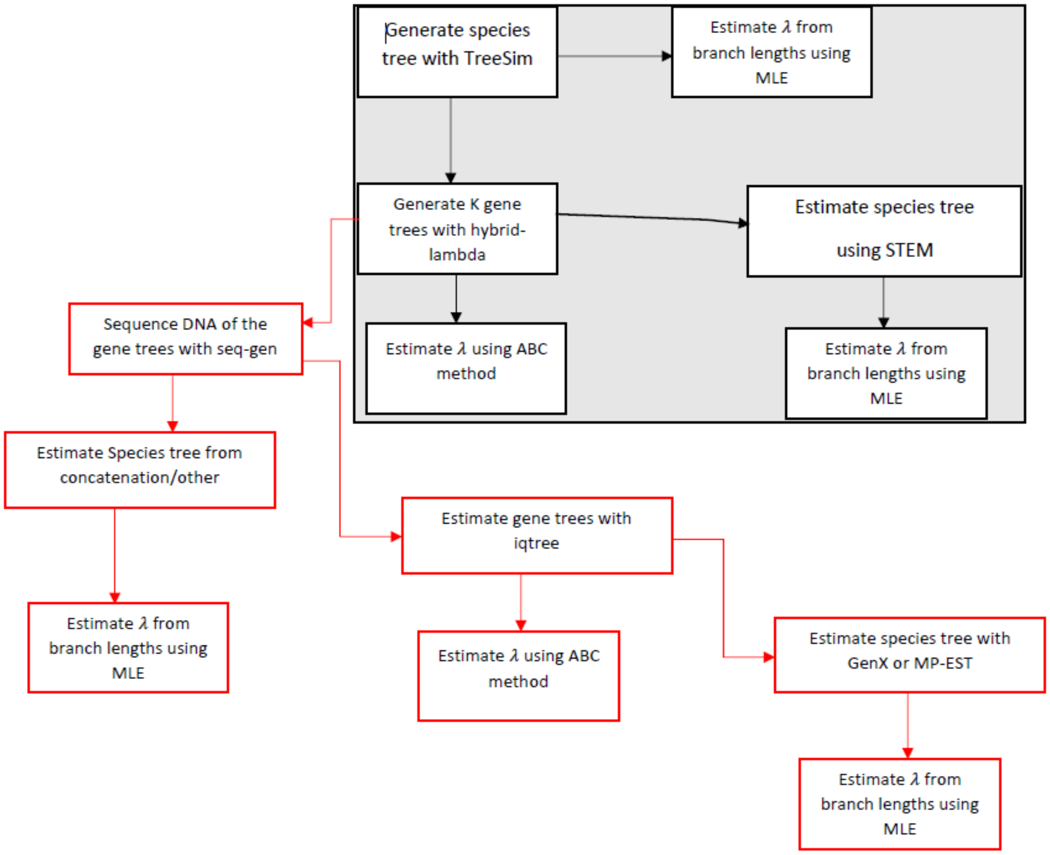
Generate species tree with TreeSim
Estimate λ from branch lengths using MLE
Generate K gene trees with hybrid-lambda
Estimate species tree using STEM
Sequence DNA of the gene trees with seq-gen
Estimate λ using ABC method
Estimate λ from branch lengths using MLE
Estimate Species tree from concatenation/other
Estimate gene trees with iqtree
Estimate λ from branch lengths using MLE
Estimate λ using ABC method
Estimate species tree with GenX or MP-EST
Estimate λ from branch lengths using MLE

Figure 4.4: Flow chart of the simulation.The shaded parts on the gray box at the top right corner are discussed in this project.

4.2.4 Result

The ABC method is used to infer the speciation rate from gene trees instead of species tree. To investigate the effect of priors, numbers of loci, numbers of replicate, and numbers of taxa of species trees on the estimates, we employed the following parameters:

- 10 taxon and 15 taxon.
- 500 and $1,000$ (loci).
- priors: exponential with means 1 and 2, and beta distribution with $\beta = \alpha = 2$.
- $10,000$ replicates of the species trees per each simulation.

Each combination of the above parameters was used and replicated 50 times. A size of $N = 10,000$ species trees for each replicate and a sample size of 500 loci were used for the each of the priors. While $10,000$ species trees, $1,000$ loci were used with exponential priors only. The 500 best speciation rates were retained. The best 500 was determined based on the smallest D_i. The value of N was determined by pilot simulation studies with $N = 50,000$ and $N = 10,000$, which suggested that $N = 10,000$ and 500 loci yielded approximate results with $N = 50,000$ and 500 loci.

Doubling the mean of the exponential prior from 0.5 to 1 widens the credibility intervals of the estimates (Tables 4.2 and 4.3). A $\boldsymbol{beta(2,2)}$ prior produces better estimates of λ and tighter credibility intervals (Table 4.4). For example, the maximum width of the credibility interval for beta distribution is 0.666 compare to 2.737 and 1.961 for ***exp*(1)** and ***exp*(2)**, respectively, for 10 taxon species, and 0.606 compare to 1.706 and 1.728 for ***exp*(1)** and ***exp*(2)**, respectively, for 15 taxon species. Further, the credibility proportion, that is the number of times the credible intervals contain the observed value ($\lambda = 0.5$) is 100% for both 10 and 15 taxa species trees for the beta prior (Table 4.1). On average, the predictive distribution of λ is not much sensitive to the shape of the priors, the numbers of taxa, and numbers of loci, particularly when the number of loci is at least 500 (Table 4.4 and Figures 4.5 and 4.6). However, the shape of priors affects the widths of the credible intervals and the proportion of times the credible intervals contain the observed value.

Prior	exp(2)	exp(1)	beta(2,2)
10-taxon			
Coverage proportion	0.960	0.920	1.000
Minimum estimate	0.243	0.262	0.306
Maximum estimate	1.325	1.665	0.744
Max. credible Interval length	1.961	2.738	0.666
15-taxon			
Coverage proportion	0.900	0.960	1.000
Minimum estimate	0.243	0.335	0.283
Maximum estimate	1.271	1.292	0.765
Max. credible Interval length	1.728	1.706	0.606

Table 4.1: Summary of Predictive Values for 10-taxon and 15-taxon

However, a prior of *exp*(2) was used for the remaining simulations based on the fact that a randomly selected tree pendant edge has an exponentially distributed length with parameter 2λ (Stadler and Steel, 2012a). Further, we examine the effect of the number of gene trees on the predictive values. Also, the sum of the branch lengths of the species trees are negatively correlated with estimated speciation rates (Figures 4.5(d,e,f) and 4.6(d,e,f)).

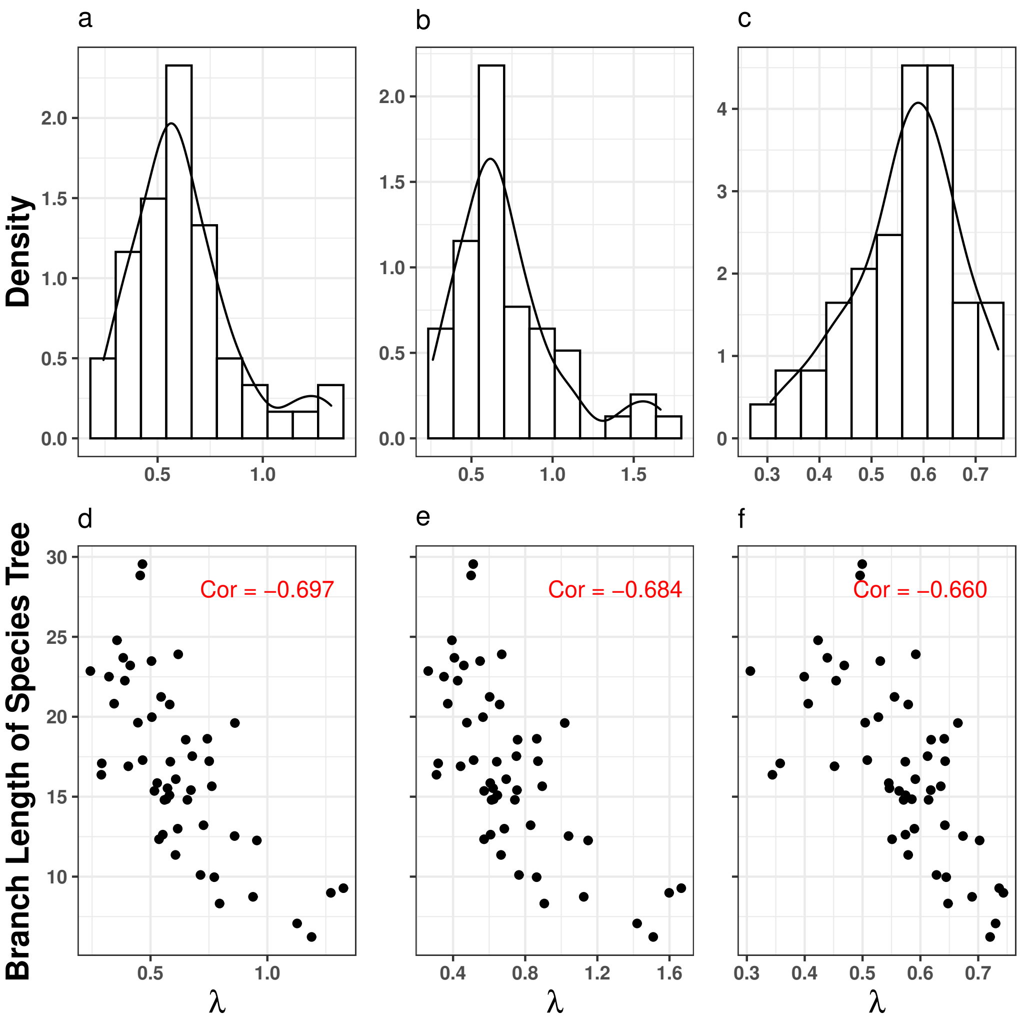

Figure 4.5: Predictive Distributional Plots of speciation rates of 10-taxon trees. (a)-(c) are the average predictive estimates for λ for exp(2), exp(1) and beta(2,2) priors, respectively. (d)-(f) are plots of speciation rate (λ) and sum of the branch lengths of the observed species tree for the three priors.

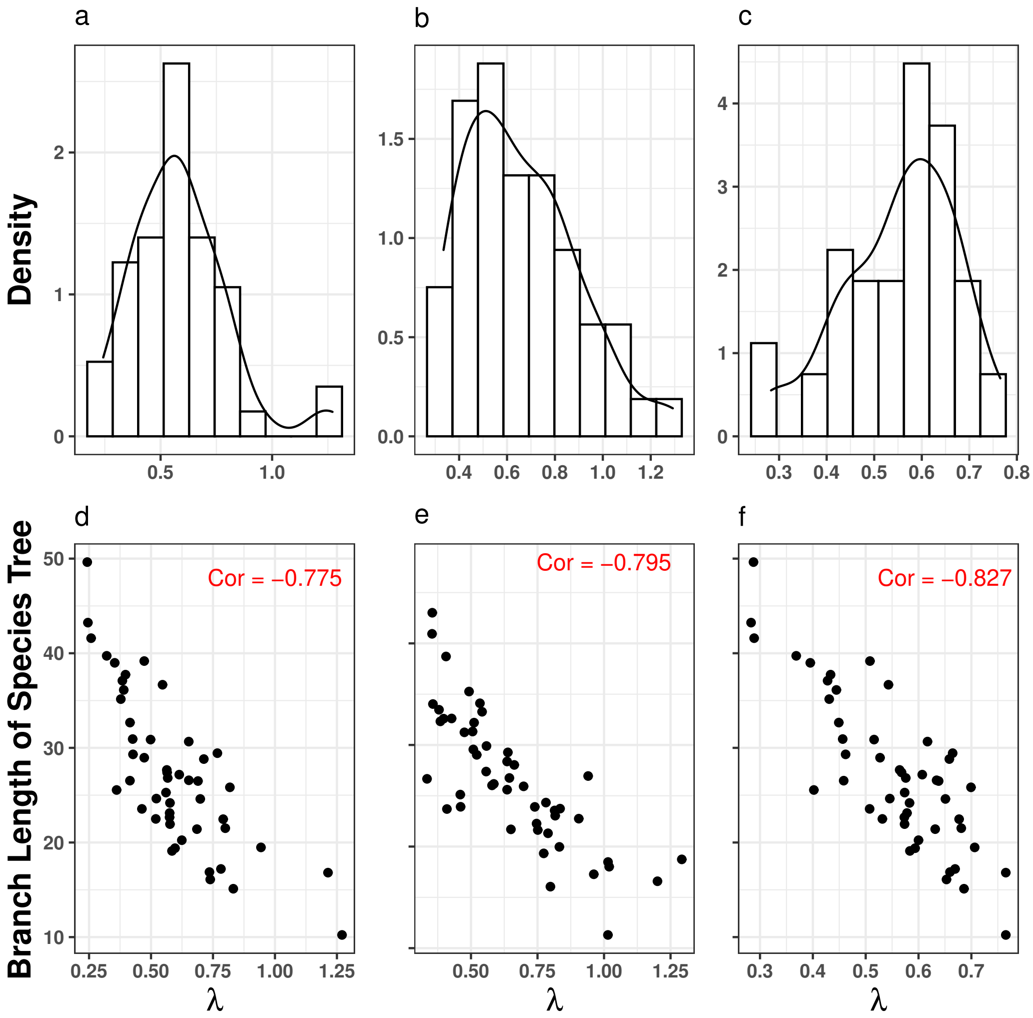

Figure 4.6: Predictive Distributional Plots of speciation rates of 15-taxon trees. (a)-(c) are the average predictive estimates for λ for exp(2), exp(1) and beta(2,2) priors, respectively. (d)-(f) are plots of speciation rate (λ) and sum of the branch lengths of the observed species tree for the three priors.

$exp(2)$			$exp(1)$		
$exp2.mean$	$quant.025$	$quant.975$	$exp1.mean$	$quant.025$	$quant.975$
0.466	0.170	0.928	0.514	0.174	1.043
1.127	0.488	2.112	1.421	0.564	2.871
0.616	0.259	1.173	0.684	0.239	1.419
0.290	0.098	0.593	0.308	0.099	0.638
0.529	0.205	1.015	0.606	0.225	1.213
0.506	0.177	0.989	0.566	0.213	1.137
0.773	0.302	1.479	0.863	0.305	1.707
0.553	0.215	1.108	0.608	0.220	1.153
0.243	0.061	0.532	0.262	0.078	0.572
0.560	0.218	1.135	0.613	0.246	1.196
0.939	0.394	1.768	1.125	0.469	2.261
0.955	0.358	1.883	1.149	0.484	2.404
0.585	0.232	1.140	0.642	0.253	1.248
0.582	0.222	1.130	0.658	0.259	1.333
1.272	0.538	2.419	1.598	0.629	3.034
0.860	0.320	1.687	1.040	0.399	2.163
0.673	0.262	1.352	0.754	0.278	1.405
0.608	0.221	1.256	0.695	0.239	1.320
0.504	0.188	0.957	0.550	0.216	1.139
0.465	0.167	0.956	0.512	0.189	1.021
0.861	0.370	1.596	1.019	0.400	1.998
0.413	0.134	0.820	0.460	0.153	0.895
0.357	0.126	0.724	0.393	0.124	0.830
0.619	0.252	1.203	0.670	0.263	1.308
0.727	0.268	1.381	0.830	0.305	1.683
0.322	0.110	0.636	0.349	0.110	0.747
0.405	0.147	0.814	0.442	0.161	0.973
0.658	0.246	1.246	0.743	0.282	1.464
0.573	0.207	1.118	0.622	0.229	1.235
0.383	0.133	0.787	0.407	0.158	0.796

Table 4.2: Estimates of λ and its credible intervals for 10 taxon species tree with $exp(2)$ and $exp(1)$ priors, respectively. $quant.025$ and $quant.975$ are the 2.5% and 97.5% quantiles of the estimated predictive distribution.

$exp(2)$			$exp(1)$		
$exp2.mean$	$quant.025$	$quant.975$	$exp1.mean$	$quant.025$	$quant.975$
0.427	0.209	0.775	0.385	0.174	0.711
0.575	0.269	1.027	0.542	0.244	0.970
0.522	0.238	0.925	0.814	0.381	1.535
0.473	0.222	0.820	0.354	0.147	0.635
0.818	0.400	1.367	1.015	0.467	1.780
0.321	0.143	0.599	0.789	0.375	1.385
0.944	0.431	1.630	0.961	0.416	1.630
0.577	0.277	1.014	0.746	0.345	1.371
1.214	0.505	2.177	0.558	0.254	1.040
1.271	0.592	2.321	0.380	0.183	0.668
0.585	0.267	1.027	0.475	0.202	0.874
0.564	0.248	1.027	0.650	0.320	1.189
0.245	0.109	0.451	0.644	0.290	1.143
0.384	0.165	0.703	0.522	0.238	0.959
0.397	0.162	0.740	0.663	0.292	1.236
0.567	0.260	0.985	0.354	0.148	0.688
0.739	0.364	1.317	0.740	0.336	1.331
0.426	0.197	0.776	0.816	0.376	1.478
0.652	0.295	1.197	0.698	0.350	1.254
0.624	0.277	1.098	0.798	0.384	1.382
0.689	0.326	1.209	0.782	0.390	1.406
0.354	0.163	0.659	1.201	0.536	2.129
0.546	0.264	0.970	1.292	0.608	2.315
0.791	0.368	1.459	0.835	0.407	1.516
0.832	0.374	1.528	0.636	0.274	1.130
0.243	0.100	0.447	0.639	0.296	1.171
0.560	0.262	0.964	0.410	0.172	0.756
0.768	0.350	1.381	0.509	0.233	0.878
0.699	0.326	1.210	0.773	0.325	1.322
0.614	0.285	1.065	0.586	0.286	1.098

Table 4.3: Estimates of λ and its credible intervals for 15 taxon species tree with $exp(2)$ and $exp(1)$ priors, respectively. $quant.025$ and $quant.975$ are the 2.5% and 97.5% quantiles of the estimated predictive distribution.

	10-taxon			**15-taxon**		
	$exp(2)$	$exp(1)$	$beta(2,2)$	$exp(2)$	$exp(1)$	$beta(2,2)$
1	0.466	0.514	0.508	0.427	0.385	0.462
2	1.127	1.421	0.730	0.575	0.542	0.573
3	0.616	0.684	0.590	0.522	0.814	0.546
4	0.290	0.308	0.344	0.473	0.354	0.508
5	0.529	0.606	0.545	0.818	1.015	0.699
6	0.506	0.566	0.527	0.321	0.789	0.369
7	0.773	0.863	0.645	0.944	0.961	0.706
8	0.553	0.608	0.574	0.577	0.746	0.583
9	0.243	0.262	0.306	1.214	0.558	0.765
10	0.560	0.613	0.571	1.271	0.380	0.765
11	0.939	1.125	0.689	0.585	0.475	0.584
12	0.955	1.149	0.702	0.564	0.650	0.564
13	0.585	0.642	0.574	0.245	0.644	0.283
14	0.582	0.658	0.579	0.384	0.522	0.428
15	1.272	1.598	0.744	0.397	0.663	0.434
16	0.860	1.040	0.674	0.567	0.354	0.576
17	0.673	0.754	0.618	0.739	0.740	0.653
18	0.608	0.695	0.591	0.426	0.816	0.456
19	0.504	0.550	0.531	0.652	0.698	0.617
20	0.465	0.512	0.499	0.624	0.798	0.600
21	0.861	1.019	0.665	0.689	0.782	0.638
22	0.413	0.460	0.468	0.354	1.201	0.395
23	0.357	0.393	0.423	0.546	1.292	0.543
24	0.619	0.670	0.592	0.791	0.835	0.677
25	0.727	0.830	0.642	0.832	0.636	0.686
26	0.322	0.349	0.399	0.243	0.639	0.288
27	0.405	0.442	0.451	0.560	0.410	0.574
28	0.658	0.743	0.614	0.768	0.509	0.665
29	0.573	0.622	0.546	0.699	0.773	0.651
30	0.383	0.407	0.439	0.614	0.586	0.607

Table 4.4: First 30 estimates of λ for 10 taxon and 15 taxon species with, 500 loci, and $exp(2)$, $exp(1)$ and $beta(2,2)$ priors, respectively.

Number of Loci

Ten and fifteen taxon species trees were generated under a pure birth model with the values of λ from the prior $exp(2)$, and 500 and 1000 loci from each species tree. Table 4.5 and Figure 4.7 show the effects of the sample size of the gene trees on the estimates of λ. The credibility intervals and coverage probability widths are approximately the same for the 500 and 1000 loci of genes for both trees. The coverage probability for 500 and 1000 loci are 0.96 for both loci for 10 taxon species tree, and 0.98 and 0.96, respectively, for 15 taxon species tree. The posterior estimates are close for these sample sizes, Table 4.5, the maximum absolute differences of the estimates using the sample size are 0.0407 and 0.026 for 10 and 15 taxon species trees, respectively. Increasing the number of loci from 500 to 1000 does not affect the estimates. However, we speculate that there would be some improvement by using a larger sample size if the gene trees were estimated instead of true gene trees. The reason is that the estimated rather than known gene trees make inference more difficult (Huelsenbeck and Kirkpatrick, 1996; Huang et al., 2010; Roch and Warnow, 2015), and by the law of large numbers, increasing the sample size is more likely to improve the estimates. Note a difference between the known gene tree versus estimated gene tree cases was that with the estimated gene trees, there are sampling and estimation errors, thus, noisy data, while the known gene trees have lesser or no noise. We have proposed to carry out the same study with estimated gene trees as part of our future projects.

Also, there is no clear difference between the shapes of the predictive distributions for the both number of loci, Figures 4.7(a) and 4.7(d) for 10 taxon tree, and Figures 4.7(e) and 4.7(h) for 15 taxon tree. In addition, some of the species' trees have shorter branch lengths. Shortness of tree branch lengths would cause more gene tree heterogeneity and discordance of gene trees and species trees, which might cause ABC to estimate a less accurate value of the observed $\lambda = 0.5$, particularly values higher than this value. Interestingly, increasing the sample size does not improve the estimates much.

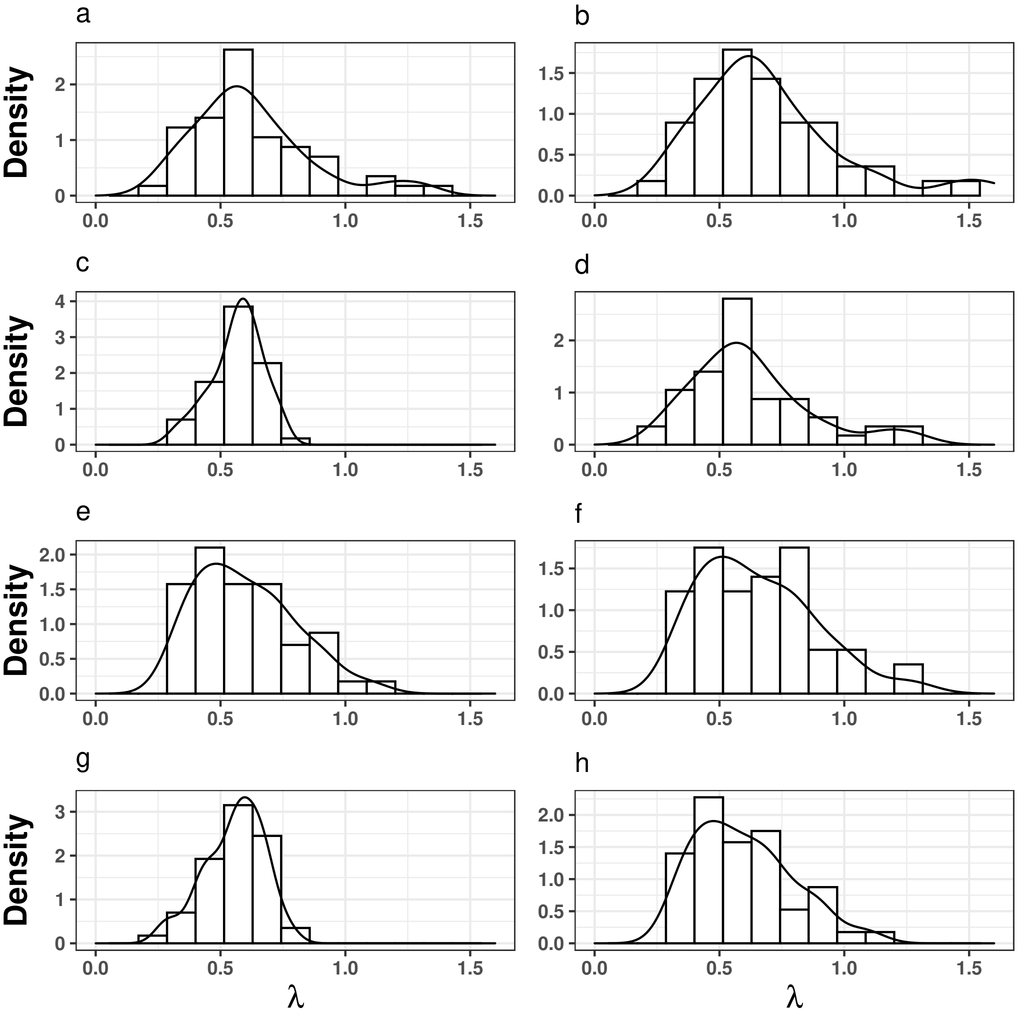

Figure 4.7: Predictive plots of speciation rates. (a)-(c) have priors: $exp(2)$, $exp(1)$ and $beta(2, 2)$, respectively, with 500 loci from 10 taxon species tree. (d) has prior of exp(2) with 1000 loci from 10 taxon species tree. (e)-(g) have priors: $exp(2)$, $exp(1)$ and $beta(2, 2)$, respectively, with 500 loci from 15 taxon species tree. (h) has prior of exp(2) with 1000 loci from 15 taxon species tree.

10-taxon			**15-taxon**		
$n_g = 500$	$n_g = 1000$	$difference$	$n_g = 500$	$n_g = 1000$	$difference$
0.466	0.472	0.005	0.368	0.363	0.005
1.127	1.141	0.014	0.497	0.487	0.010
0.616	0.624	0.008	0.759	0.737	0.022
0.290	0.280	0.010	0.339	0.339	0.000
0.529	0.542	0.014	0.906	0.894	0.012
0.506	0.499	0.007	0.711	0.701	0.011
0.773	0.786	0.013	0.903	0.893	0.009
0.553	0.551	0.001	0.672	0.668	0.003
0.243	0.243	0.000	0.530	0.536	0.005
0.560	0.570	0.010	0.368	0.368	0.000
0.939	0.912	0.026	0.452	0.457	0.005
0.955	0.991	0.036	0.607	0.610	0.003
0.585	0.583	0.001	0.587	0.588	0.001
0.582	0.574	0.008	0.465	0.477	0.012
1.272	1.242	0.030	0.597	0.602	0.005
0.860	0.900	0.041	0.336	0.337	0.001
0.673	0.670	0.003	0.680	0.666	0.014
0.608	0.607	0.001	0.740	0.740	0.000
0.504	0.491	0.013	0.651	0.650	0.000
0.465	0.457	0.008	0.719	0.708	0.012
0.861	0.884	0.023	0.695	0.697	0.002
0.413	0.418	0.005	1.034	1.051	0.018
0.357	0.362	0.005	1.113	1.103	0.011
0.619	0.626	0.007	0.761	0.779	0.018
0.727	0.731	0.005	0.583	0.597	0.014
0.322	0.321	0.001	0.585	0.575	0.010
0.405	0.418	0.013	0.389	0.401	0.011
0.658	0.658	0.001	0.469	0.477	0.008
0.573	0.581	0.008	0.702	0.692	0.010
0.383	0.388	0.005	0.561	0.569	0.008

Table 4.5: Estimates of speciation rates with 500 and 1000 loci and their differences for 10 and 15 taxon species trees.

4.2.5 Comparison of Results: Maximum Likelihood Method versus ABC Method

Tables 4.7 and 4.8 summarize the estimated values of λ for 10-taxon and 15-taxon species trees, respectively, for using the maximum likelihood (MLE) method on the branch lengths of known species trees, estimated species trees, and ABC method. The summary statistics for the estimates in Table 4.6. The means of the estimates are 0.581, 0.580, and 0.580 for using the MLE method on the branch lengths of known species trees and estimated species trees from 500 and $1,000$ loci, respectively; also, 0.626 and 0.627 for the ABC method with 500 and 1000 loci, respectively, for 10-taxon species trees. Further, similar results were obtained for 15-taxon species trees (Table 4.6 and Figure 4.8. The estimates with the ABC method are a little higher than those with the maximum likelihood method. We speculate that the ABC method would produce approximate results by increasing the number of replicates and loci.

Statistic	Known ST	Est_g500	Est_g1000	ABC_g500	ABC_g1000
10-taxon					
min	0.250	0.250	0.250	0.243	0.243
median	0.528	0.528	0.528	0.582	0.582
mean	0.581	0.580	0.580	0.626	0.627
stdv	0.214	0.214	0.214	0.245	0.243
max	1.217	1.216	1.217	1.325	1.302
15-taxon					
min	0.294	0.294	0.294	0.307	0.313
median	0.534	0.534	0.534	0.583	0.576
mean	0.561	0.560	0.561	0.600	0.599
stdv	0.165	0.164	0.164	0.196	0.195
max	0.986	0.985	0.986	1.113	1.103

Table 4.6: Summary statistics for the estimates of λ for the 10-taxon and 15-taxon species trees, from the branch lengths of known and estimated species trees, and ABC method with 500 and 1000 loci.

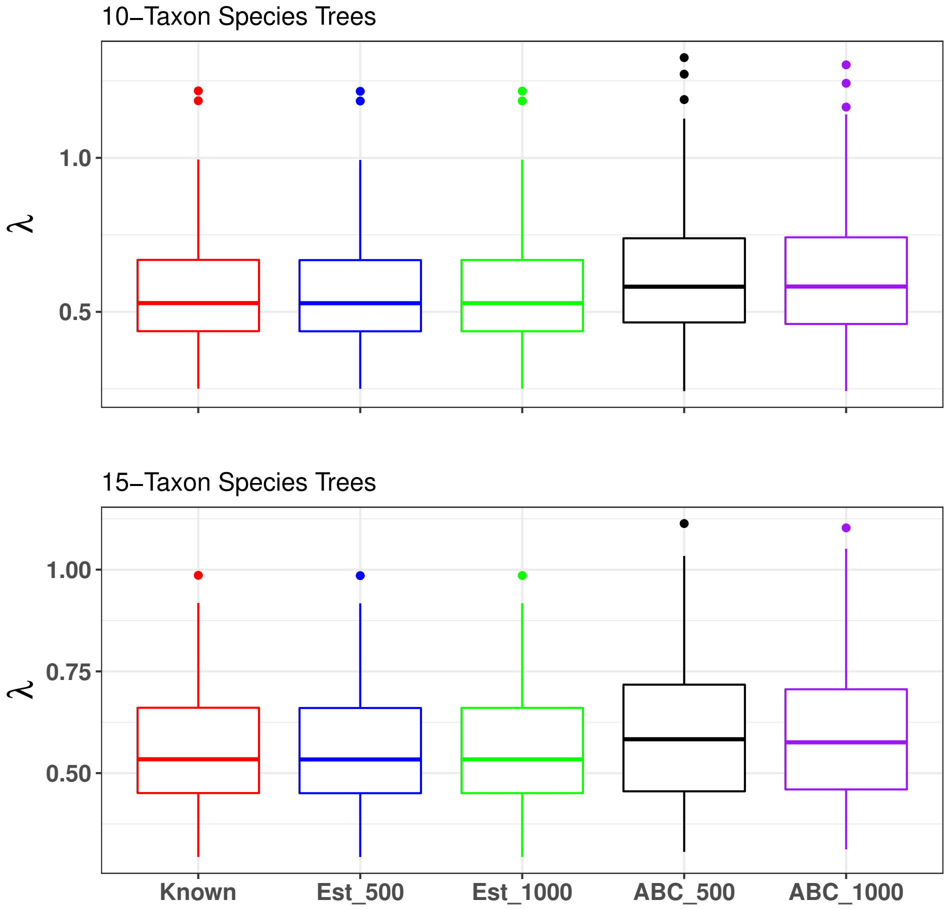

Figure 4.8: Boxplots for the estimated values of λ from the branch lengths of known species trees, estimated species trees and ABC method with 500 and $1,000$ loci.

	True ST	ST_g500	ST_g1000	ABC_g500	ABC_g1000
1	0.426	0.426	0.426	0.466	0.472
2	0.808	0.807	0.807	1.127	1.141
3	0.544	0.543	0.543	0.616	0.624
4	0.977	0.976	0.976	0.290	0.280
5	0.472	0.472	0.472	0.529	0.542
6	0.724	0.724	0.724	0.506	0.499
7	0.994	0.993	0.994	0.773	0.786
8	0.250	0.250	0.250	0.553	0.551
9	0.277	0.277	0.277	0.243	0.243
10	0.501	0.500	0.500	0.560	0.570
11	0.549	0.549	0.549	0.939	0.912
12	0.297	0.297	0.297	0.955	0.991
13	0.833	0.831	0.832	0.585	0.583
14	0.674	0.674	0.674	0.582	0.574
15	0.629	0.629	0.629	1.272	1.242
16	1.185	1.185	1.185	0.860	0.900
17	0.732	0.732	0.732	0.673	0.670
18	0.511	0.511	0.511	0.608	0.607
19	0.511	0.511	0.511	0.504	0.491
20	0.424	0.424	0.424	0.465	0.457
21	0.803	0.802	0.803	0.861	0.884
22	0.708	0.707	0.707	0.413	0.418
23	0.585	0.584	0.585	0.357	0.362
24	0.487	0.486	0.486	0.619	0.626
25	0.513	0.513	0.513	0.727	0.731
26	0.373	0.373	0.373	0.322	0.321
27	0.292	0.292	0.292	0.405	0.418
28	0.404	0.404	0.404	0.658	0.658
29	0.440	0.439	0.440	0.573	0.581
30	0.460	0.459	0.459	0.383	0.388

Table 4.7: The first 30 estimates of λ for 10-taxon species (known) trees. Column 1 is the estimates directly from the branch lengths of the species tree. Column 2 is the estimates from the branch lengths of the species tree estimated with 500 gene trees. Column 3 is the estimates from the branch lengths of the species tree estimated with 1000 gene trees. Column 4 is the estimates with ABC method from 500 gene trees. Column 5 is the estimates with ABC method from 1000 gene trees

	True ST	ST_g500	ST_g1000	ABC_g500	ABC_g1000
1	0.492	0.492	0.492	0.368	0.363
2	0.757	0.757	0.757	0.497	0.487
3	0.442	0.442	0.442	0.759	0.737
4	0.918	0.917	0.918	0.339	0.339
5	0.574	0.574	0.574	0.906	0.894
6	0.874	0.873	0.873	0.711	0.701
7	0.847	0.846	0.846	0.903	0.893
8	0.294	0.294	0.294	0.672	0.668
9	0.327	0.327	0.327	0.530	0.536
10	0.390	0.390	0.390	0.368	0.368
11	0.580	0.579	0.580	0.452	0.457
12	0.366	0.366	0.366	0.607	0.610
13	0.986	0.985	0.986	0.587	0.588
14	0.540	0.539	0.540	0.465	0.477
15	0.729	0.729	0.729	0.597	0.602
16	0.696	0.695	0.695	0.336	0.337
17	0.649	0.648	0.648	0.680	0.666
18	0.397	0.397	0.397	0.740	0.740
19	0.625	0.625	0.625	0.651	0.650
20	0.372	0.372	0.372	0.719	0.708
21	0.597	0.596	0.596	0.695	0.697
22	0.709	0.709	0.709	1.034	1.051
23	0.643	0.642	0.642	1.113	1.103
24	0.499	0.498	0.499	0.761	0.779
25	0.459	0.459	0.459	0.583	0.597
26	0.464	0.464	0.464	0.585	0.575
27	0.304	0.304	0.304	0.389	0.401
28	0.457	0.456	0.456	0.469	0.477
29	0.534	0.534	0.534	0.702	0.692
30	0.465	0.465	0.465	0.561	0.569

Table 4.8: The first 30 estimates of λ for 15-taxon species (known) trees. Column 1 is the estimates directly from the branch lengths of the species tree. Column 2 is the estimates from the branch lengths of the species tree estimated with 500 gene trees. Column 3 is the estimates from the branch lengths of the species tree estimated with 1000 gene trees. Column 4 is the estimates with ABC method from 500 gene trees. Column 5 is the estimates with ABC method from 1000 gene trees

4.3 Discussion

To our knowledge, our method is the first method we are of to estimate the speciation rate from gene trees under a pure birth model. Within the past few decades, some version of the random Markov chain Montel Carlo (MCMC) sampling method and Snyder filter (SF) (Parag and Pybus, 2018), a likelihood under sparse sampling scenario (Kayondo et al., 2019), multitype birth-death model (MTBD) in BEAST 2 (Barido-Sottani et al., 2020) were developed to infer the speciation rate either from species tree or allele frequencies but not from gene trees. In practice, species trees are readily available. Species trees are inferred from the gene trees, which have also been estimated from DNA sequences. Note that our focus has not been on comparing the ABC method to some of these methods in the project, but we did small simulations with the likelihood method, STEM (Kubatko et al., 2009), to check its performance (Tables 4.7 and 4.8, and Figure 4.8).

The ABC method uses summary statistics, which are sometimes not sufficient, and might not guarantee that the estimated predictive (posterior) distribution converges to the true posterior distribution (Marjoram and Tavaré, 2006). However, it is still common to use the ABC method in phylogenetic and genetic studies (Tanaka et al., 2006; Stadler, 2011b), and in our case, we note that the pairwise RF distance is not a sufficient statistic. Recall, for a summary statistics T to be sufficient statistics, $T(\boldsymbol{X}) = T(\boldsymbol{Y})$ for any given two data sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ (Casella and Berger, 2021). However, two data sets each with different priors ($exp(2)$ and $exp(1)$) indicate that best βN simulated RF distances with the smallest absolute distance to the observed RF distance approximately yield the same predictive distributions, and about 95% of the credible intervals contain the true parameter. However, pairwise RF distances still contain information about the speciation rate of the tree. Similar results for 10 and 15 taxon species trees with different priors and number of loci illustrate very little information loss in using this approach, despite the lack of sufficiency.

4.4 Computation Time

The ABC method for estimating the speciation rate is slow due to the number of computations. On average, it scales well with the number of taxa; we infer that the average time required to run the above algorithm for one replicate, which consists of $10,000$ species trees and 500 loci per tree, were 4.34 and 6.56 hours for 10 and 15 taxa, respectively. Also, 8.6 and 11.6 hours were used with $1,000$ loci per species tree for 10 and 15 taxa, respectively. This suggests that the computation time is roughly linearly dependent on the number of taxa. The computation scaled well and was faster with the parallel framework.

Further, it would be better to increase N for more taxa since there are fewer speciation rates simulated in the prior(s) that match the true speciation rate ($\lambda = 0.5$). Alternative approaches for inferring the birth-death parameters (Tanaka et al., 2006; Stadler, 2011b; Barido-Sottani et al., 2020) use maximum likelihood to infer these parameters from the branching structures of the species trees. These methods have not been implemented with gene trees; however, likelihood calculations with gene trees scale slowly in the number of taxa (Rosenberg, 2007; Disanto and Rosenberg, 2015; Truszkowski et al., 2021), indicating that this method will not scale well in the number of taxa. The available techniques calculate the product of the likelihood of the lineages evolving through different tree branches from the tips; thus, calculating these likelihoods for at least 10-taxon species trees requires calculating the likelihood for up to N possible gene trees since there is a reasonable chance that all gene tree topologies are unique (chapter 2). The advantage of the ABC method is that it does not depend on calculating the likelihoods, simulating gene trees from species trees, and computing the RF distances scale well with the number of taxa. However, we speculate that when the estimated gene trees are used, it is likely to take longer due to additional time to infer the gene trees from the DNA sequence.

4.5 Application to Empirical Datasets

Application to Empirical Gibbons Dataset

We randomly sampled a set of 1000 gene trees from 10706 set of the Gibbons dataset with five taxa: *Hylobates moloch* (HMO), *Hylobates pileatus* (HPL), *Nomascus leucogenys* (NLE), *Hoolock leuconedys* (HLE) and *Symphalangus syndactylus* (SSY), and used it to infer the speciation rate of the organisms using ABC method. The species HMO and HPL are the genus *(*Hylobates) (H), and NLE, HLE and SSY represent the genera *Nomascus* (N), *Hoolock* (B) and *Symphalangus* (S), respectively. These gene trees were estimated from the gibbon sequence dataset (Kim and Degnan, 2020).

The speciation rate λ was estimated from the 1000 sampled gene trees with prior of the $exp(1)$. The estimate is 0.88 with 95% credibility interval of $(0.208, 2.424)$, (see Figure 4.9).

4.6 Conclusion

This study provides an alternative method to infer the birth-death parameters from gene trees with some number of loci. The ability of ABC to infer the speciation rate seems not to be sensitive to the priors and the numbers of loci ($n_g \geq 500$). We note that more work needs to be done to see the effect of using larger numbers of loci, species trees, and replicates. Sequence data is not needed for the ABC method since gene trees are its only data source. Using summary statistics instead of sufficient statistics does not show much loss of information in the data set. However, larger numbers of loci and replicates might increase the accuracy and decrease the variability of the estimates. The ABC method could be used with a diffuse prior, or a more informative prior. An informative prior was used in this study. The possibility that under typical birth-death processes, some trees are more probable than others for larger numbers of taxa and the prior could be based on this rather

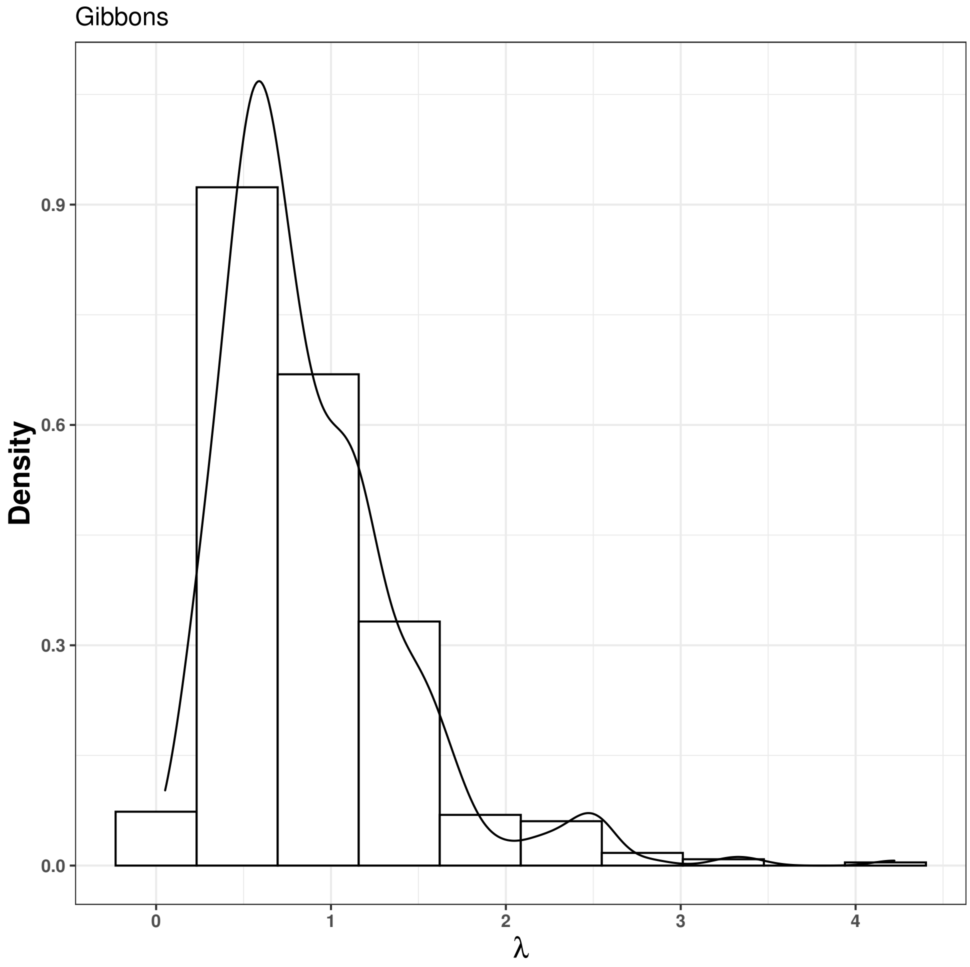

Figure 4.9: Distributional plot for the estimated values of λ from the Gibbons gene trees.

than making each speciation rate equally likely in the prior.

In the 10 and 15 taxa cases, about 95% of the credibility regions contain the observed value. We propose to use empirical data and other simulation settings outlined on Figure 4.4 such as estimated gene trees with and without mutations to compare the performance of the ABC method in future studies.

Chapter 5

Conclusions and Future Works

5.1 Conclusion

A phylogenetic tree defines a graphical framework for modeling relationships between species. Phylogenetic trees are hypotheses, and not definitive facts. The branching structure in a phylogenetic tree reflects how species or other groups evolved from a series of common ancestors in the past. A species tree under the multispecies model defines a unique distribution of gene trees. The development of statistical methods for inference of species trees from gene trees or DNA sequences has been on the rise for over two decades. Many of these methods have been found to be statistically consistent under some criteria (Kubatko et al., 2009; Liu et al., 2009; Liu and Edwards, 2009; Mirarab et al., 2014).

Despite the availability of these methods, many empirical results show much variability among the gene trees or between gene trees and species trees, and this discord is often reported to be surprising (Salichos and Rokas, 2013), and making difficult to estimate the true tree.

Chapter one found the bounds of the probability of uniqueness of gene tree topologies for a given species tree and the number of loci, holding other biological factors constant. These results hold for either unranked rooted or ranked rooted gene trees, but the bounds would be wider for the latter case. However, we focus on unranked rooted binary trees since, in practice, preservation of the order of the nodes of the gene tree together with their topological relationships among gene lineages is less important. The bounds show that it is necessary to take into account the number of taxa of the species and number of loci when checking the level of heterogeneity of the gene tree topologies, particularly for a large number of species (i.e., more than ten taxa). Furthermore, the growth of the possible number of unranked or unranked rooted binary gene trees for n taxon species tree is faster than exponential for large n. So the fact that the topologies of the sampled gene trees can be distinct might not hold for a small number of loci and a large number of taxa.

Simulation studies in chapter 2 confirmed the results in chapter 1, that numbers of taxa and loci contribute to the variability of gene tree topologies, especially, when the speciation rate is high. Higher speciation rates lead to shorter branches in the species tree and, therefore, higher incomplete lineage sorting (deep coalesce) (Harvey and Rambaut, 1998;

Paradis, 2016; Kim and Degnan, 2021). We observed that the probability of all topologies being distinct increases with the number of taxa and speciation rates and rapidly for $\mu = 0$. Both results are intuitive: for increasing numbers of taxa, there are more possible gene trees from the species tree, and increasing the speciation rate shortens the branch lengths and increases the number of the tree branches and denseness of the tree structure. So, on average, a species tree with short branches has a high diversity rate $(\frac{\lambda}{\mu})$ and is likely to produce diverse and distinct gene tree topologies. Further, the probability of uniqueness decreases as the sample size increases.

Chapter three discussed the ABC method for inferring the speciation rate for a pure birth model from the gene trees. The method estimates the posterior distribution of the parameter from the best βN values of the prior that produced the simulated data (sum of pairwise RF distances of the gene trees sampled from the species tree) close to the observed data. The motivation is from the correlation of the pairwise RF distance among the gene trees and the RF distance between the gene trees and the species tree. The result from this method indicates that it is possible to estimate the birth parameter from gene trees, but we note that more extensive studies are needed to check the performance and robustness of the method.

5.2 Future Works

The main shortcoming of the ABC method is that it uses summary statistics instead of sufficient statistics. Also, it needs a large simulation study to confirm its accuracy and robustness. Possible areas for future studies as depicted on the Figure 4.4 are:

- use the likelihood-based and pseudo-likelihood methods to infer the birth-death parameters and compare the results.
- use the estimated gene trees instead true trees as input data in the ABC and the likelihood-based methods to infer the birth-death parameters and compare the results.
- apply the ABC method to empirical data and compares the results with the likelihood-based methods.

The classical representation of evolution (phylogenetic tree) of a set of species often fails to represent the relationship among the species due to some processes such as recombination, hybridization, and horizontal gene transfer. Phylogenetic networks are used to depict complex relationships of a set of species. Networks capture the implicit inheritance of genetic material through gene flow and are more general for modeling the evolutionary history of species (Solís-Lemus and Ané, 2016; Degnan, 2018). We have proposed to extend the simulation studies in Chapter 2 to phylogenetic network to investigate the factors responsible for the heterogeneity of the gene tree topologies.

Chapter 6

Appendix

A: Proof of Lemma 2.2.2

Lemma 6.0.1. *For any* $k \geq 0$, $\sum_{i=1}^{m-1} i^k \leq \frac{m^{k+1}}{k+1}$

Proof.

Let $y \geq 0$. Then, by binomial expansion,

$$\int_i^{i+1} y^k dy = \frac{y^{k+1}}{k+1}|_i^{i+1} = i^k + \frac{k}{2}i^{k-1} + ... + \frac{1}{k+1}$$

$$\therefore \quad i^k \leq \int_i^{i+1} y^k dy \;\; \forall \;\; y \geq 0$$

Summing over $i = 0, 1, ..., m-1$, we have,

$$\sum_{i=0}^{m-1} i^k \leq \int_0^{m-1} y^k dy \leq \int_0^m y^k dy = \frac{y^{k+1}}{k+1}|_0^m = \frac{m^{k+1}}{k+1} \tag{6.1}$$

B: Scripts for Simulations in Chapter 2

Here we provide some representative scripts that were used for the simulations at various stages of this project. A few trivial scripts were selected to display. Most of the simulations were run on the supercomputer at the UNM Center for Advanced Research Computing lab.

R codes and Linux scripts

Several scripts are used to simulate species trees from a birth-death model and generate gene trees from the species tree. Also, the scripts are used to compute an RF distance among the gene trees and between the gene and species trees; the number of matching topologies and the probability of every gene tree topology is unique.

```
                 ##########################
                 ##    tree_sim.r        ##
                 ##########################
## This is an R script that generates the species trees.
## load libraries
library(ape)
library(geiger)
library(TreeSim)
set.seed(151515+000000+202020+100)
# Note:20 for lamda, 20 for n,0 for mu and 100 for
#numbsim = 100.

x <- sim.bd.taxa(n=15,numbsim=100,lambda=.1,mu=0.05,
complete= FALSE)
for (i in 1:length(x)){write.tree(x[[i]], "st15_all.txt",
append=TRUE)
```

```
                    ##########################
                    ##     add.outgroup.r     ##
                    ##########################

## This is an R script that roots the tree and drop the
# outgroup.


## load libraries
library(ape)
tree1 <- read.tree("st_temp.txt")
tree1string <- write.tree(tree1)
tree1string <- gsub(";","",tree1string)  # remove semi-colon
tree1string <- paste("(",tree1string,":10.0,t_out:1.0);"
,sep="") # add outgroup 't_out', added tip
yt1 <- read.tree(text=tree1string)
#Note, yt1 is NOT ultrametric
library(maps)
library(phytools)
tree1_ultra <- force.ultrametric(yt1)
# makes the tree ultrametric
```

```
write.tree(tree1_ultra, "st_temp2.txt")
```

```
                      ##########################
                      ##     Root_and_drop.r    ##
                      ##########################
## This is an R script that roots the tree and drop the
#outgroup.
## load required libraries.
library(ape)
library(maps)
library(phytools)

a <- read.tree("gg_temp.txt")## reads in gene trees

lapply((lapply(lapply(a,root,outgroup="t_out_1"),drop.tip,
tip="t_out_1")),
write.tree,"gt_temp2.txt",append=TRUE)
# roots, drops the outgroups and writes out the trees

d <- read.tree("gt_temp2.txt") ## reads in the trees.
f <- lapply(d, force.ultrametric)
```

```
## makes the trees to ultrametric

## loops over the ultrametric gene trees and write them to

# a file

for(i in 1:length(f)){

write.tree(f[[i]], "gt_temp3.txt", append=TRUE)

}

b <- read.tree("st_temp2.txt")## reads in a species tree

## drop the outgroup from species tree and write the tree out

write.tree(force.ultrametric(drop.tip(b, tip="t_out")),

"st_temp3.txt")


                      ###########################

                      ##     all.summary.r     ##

                      ###########################

## This is an R script that summarizes datasets from the

## simulations

## load libraries

library(gtools) # for sorting files

library(dplyr)

# function to get the mean of the RF distances

get_meanRF <- function(file_name){
  file <- read.table(file_name, header = FALSE)
  RF <- mean(file[,3])
  return(RF)
}
# function to count the numbers of matching topologies
get_matchingnumber <- function(file_name){
  file <- read.table(file_name, header = FALSE)[,3]
  match <- length(file[file==0])/2
  return(match)
}
## read in the files and summarize them on the fly
#library(gtools) # for sorting files

RF_to_ST <- NULL; nUnique <- NULL; pairwiseRF.j <- NULL
numbermatching <- NULL
for(j in 1:5){
datf1.j <- list.files(pattern = "1f-st15.txt",
full.names = FALSE)
nUnique.j <- read.table(datf1.j)[,1]
datf2.j <- list.files(pattern = "1RF_ST15",
full.names = FALSE)

datf2.j <- mixedsort(datf2.j)

datf3.j <- list.files(pattern = "1pairrwiseRF15",

full.names = FALSE)

datf3.j <- mixedsort(datf3.j)

RF_to_ST[j] <- unlist(lapply(datf2.j, get_meanRF))

pairwiseRF[j] <- unlist(lapply(datf3.j, get_meanRF))

numbermatching[j] <- unlist(lapply(datf2.j,

get_matchingnumber))

nUnique[j] <- ifelse(nUnique.j < 1000,  0, 1)

}

lambda <- rep(.1,100)

mu<- rep(0,100)

replicate <- seq(1,100)

n <- rep(15,100)

ngenetrees <- rep(1000,100)


# dataframe

df1 <- data.frame(n, lambda, mu, ngenetrees,

replicate, nUnique,RF_to_ST,pairwiseRF,numbermatching)
```

```
write.csv(df1, "sdf_iqtree15_.1_0.csv", row.names = FALSE)


                 ########################
                 ##    seq_iqtree.bash   ##
                 ########################
## This script calls all the above scripts to accomplish its
#job.
#!/bin/bash
# Load required modules
module load miniconda3-4.8.2-gcc-10.2.0-zu7qwdd
module load parallel-20200822-gcc-10.2.0-v75uus5
module load r-4.0.4-gcc-10.2.0-sancozx
source activate phylo_tools
module load iq-tree-1.7-beta12-gcc-7.4.0-e4x7qjf
cd  $PBS_O_WORKDIR
# An R script that generate the species trees.
Rscript tree_sim.r

filename="species_tree"
```

```
C=1
while read line;
do
echo "$line" > st_temp.txt
# an R script that adds outgroup to st_temp.txt.
Rscript add.outgroup.r
rm -f st_temp.txt # deletes the file on the fly.
# run hybrid-Lambda on st_temp2.txt to generate gene trees
#‘OUT_coal_unit’
hybrid-Lambda -spcu st_temp2.txt -num 1000 -seed
11\$C\99$C\12$C

## Parallel setup to speed up the simulations and computations
dir=$PBS_O_WORKDIR

parallel --wd $PBS_O_WORKDIR --env PATH --delay 1
--sshloginfile $PBS_NODEFILE \
  ‘head -{} OUT_coal\_unit | tail -1 > gt_temp\_{}$
  # grab each line of temp_gt.txt and run seq-gen on it to
  #generate dna-seq
  seq-gen -l500 -s.001 -mGTR -a1.0 -g4 -i.1
  -f.4,.1,.2,.3 -z2{}{}\01{}\5{}5{} -op < gt_temp_{}
  > dna-seq{}

#move dna-seq to iqtree to reconstruct/estimate tree from DNA
# sequence
  iqtree -s dna-seq{} -nt 1 -m GTR+I+G
  rm -f gt_temp_{}` ::: {1..1000}

for i in {1..1000}; do
        cat dna-seq${i}.treefile >> gg_temp.txt
done
rm -f dna-seq*

Rscript root_and_drop.r # an R script that roots the trees and
#drops
rm -f gg_temp.txt
rm -f gt_temp2.txt
sed 's/_1//g' gt_temp3.txt > temp_gt.txt # for removing '_1'
# from the file
```

```
## Relabel the tips of the trees to avoid the segmentation
#problems
##when using PRANC to count the number of unique gene tree
#topologies

sed -e 's/t15/o/g' -e 's/t14/n/g' -e 's/t13/m/g'
-e 's/t12/l/g' -e 's/t11/k/g' -e 's/t10/j/g' -e 's/t9/i/g'
-e 's/t8/h/g' -e 's/t7/x/g' -e 's/t6/f/g' -e 's/t5/y/g'
-e 's/t4/d/g' -e 's/t3/c/g'-e 's/t2/b/g' -e 's/t1/a/g'
temp_gt.txt > gt_temp4.txt

## remove branch lengths and Run pranc on gt_temp3 and count
# number of unique topologies
sed -re 's/[0-9]*\.[0-9]*//g' -e 's/:+//g' -e 's/[e-]//g'
-e 's/[0-9]//g' < gt_temp4.txt > gt_temp5.txt

## Run pranc
pranc -utopo gt_temp5.txt
wc -l outFreqs.txt >> 1f-st15.txt
```

```
# Calculate RF-distance
mv temp_gt.txt intree ## for pairwiseRF distances
mv st_temp3.txt intree2 ## for RF_ST distances

rm -f gt_temp4.txt gt_temp5.txt gt_temp3.txt
## Run treedist
~/treedist << EOF
D
R
2
P
S
Y
EOF
cp outfile 1pairrwiseRF15_$C.txt
# file for pairwise RF-distance between gene trees
rm -f outfile

~/treedist << EOF
```

```
D

R

2

L

S

Y

EOF

cp outfile 1RF_ST15_$C.txt

rm -f outfile

rm -f st_temp2.txt


let C=C+1

done < $filename

## an R script that summarizes the outcome of the simulation.

Rscript all.summary.r

echo Well done
```

C: ABC Method

An R script that generate the species trees from TreeSim is:

```
library(TreeSim)
## This script generates files of species trees for
## simulations of the gene trees with hybrid-Lambda
set.seed(221204)
for(j in 1:50){
 # set.seed(221204 + (j-1))

  x <- rexp(10000,2)
  write(x,file=paste("L",j,".txt",sep=""), ncol=1)
  for(i in 1:10000){
    y <- sim.bd.taxa(10,1,x[i],0)
write.tree(y[[1]],file=paste("ST",j,".txt",sep=""),append=T)
  }
}
```

```
### observed trees
set.seed(221204)
w <- sim.bd.taxa(10,50,0.5,0)
for(i in 1:50){
write.tree(w[[i]],file="STtrue.txt",append=T)
}
```

```
Bash script for ABC simulation and computation
```

```
#!/bin/bash
## simulate data to compare with observed data

Rscript abc_trees.txt # R script for generating species trees
for ((j=46 ; j<=50 ; j++))
do

cat ../ST$j.txt > file.txt

for ((c=1 ; c<=10000 ; c++))
```

```
do
head -$c file.txt | tail -1 > sim.txt

## run hybrid-Lambda on st_temp2.txt to generate OUT_coal_unit

~/bin/hybrid-Lambda -spcu sim.txt -num 500 -seed 12$c\12$c

# Removing ’_1’ from the file
sed ’s/_1//g’ OUT_coal_unit > intree

# calculate pairwise RF-distance
~/treedist.txt << EOF
D
R
2
P
S
Y
EOF
```

```
## Calculate the average of the pairwiseRFs

awk‘{total+=$3;c++}END{print total}‘outfile | cat >> RF$j.txt

rm -f outfile
rm -f intree
done
done
echo Well done
```

References

A. R. A. Alanzi and J. H. Degnan. Inferring rooted species trees from unrooted gene trees using approximate bayesian computation. *Molecular Phylogenetics and Evolution*, 116: 13–24, 2017.

S. Ali. *The book of Indian birds 11th ed.* Oxford, 1990.

M. D Anderson and T. A. Anderson. A breeding island for lesser flamingos phoeniconaias minor at kamfers dam, kimberley, south africa. *Bulletin of the African Bird Club*, 17(2): 225–228, 2010.

H. Andersson and T. Britton. *Stochastic epidemic models and their statistical analysis*, volume 151. Springer Science & Business Media, 2012.

C. Ané, B. Larget, D. A. Baum, S. D Smith, and A. Rokas. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution*, 24(2):412–426, 2007.

Mukul S Bansal, Yi-Chieh Wu, Eric J Alm, and Manolis Kellis. Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics*, 31(8):1211–1218, 2015.

J. Barido-Sottani, T. G. Vaughan, and T. Stadler. A multitype birth–death model for bayesian inference of lineage-specific birth and death rates. *Systematic Biology*, 69(5): 973–986, 2020.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

M. C. Borja. A variation of the birthday problem, 2016.

S. Bossert, E. A. Murray, A. Pauly, K. Chernyshov, S. G. Brady, and B. N. Danforth. Gene tree estimation error with ultraconserved elements: An empirical study on pseudapis bees. *Systematic Biology*, 70(4):803–821, 2021.

D. Bryant, R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A RoyChoudhury. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, 29(8):1917–1932, 2012.

L. Cai, Z. Xi, E. M. Lemmon, A. R. Lemmon, A. Mast, C. E. Buddenhagen, L. Liu, and C. C. Davis. The perfect storm: gene tree estimation error, incomplete lineage sorting, and ancient gene flow explain the most recalcitrant ancient angiosperm clade, malpighiales. *Systematic Biology*, 70(3):491–507, 2021.

REFERENCES

M. Camarri and J. Pitman. Limit distributions and random trees derived from the birthday problem with unequal probabilities. *Electronic Journal of Probability*, 5:1–18, 2000.

G. Casella and R. L Berger. *Statistical inference*. Cengage Learning, 2021.

D. Chesters and A. P. Vogler. Resolving ambiguity of species limits and concatenation in multilocus sequence data for the construction of phylogenetic supermatrices. *Systematic Biology*, 62(3):456–466, 2013.

J. Chifman and L. Kubatko. Quartet inference from snp data under the coalescent model. *Bioinformatics*, 30(23):3317–3324, 2014.

J. A. Coyne and et al Orr, H. A. *Speciation*, volume 37. Sinauer Associates Sunderland, MA, 2004.

K. A. Cranston, B. Hurwitz, D. Ware, L. Stein, and R. A. Wing. Species trees from highly incongruent gene trees in rice. *Systematic Biology*, 58:489–500, 2009.

K. Csilléry, M. G. B. Blum, O. E. Gaggiotti, and O. François. Approximate bayesian computation (abc) in practice. *Trends in Ecology & Evolution*, 25(7):410–418, 2010.

A. DasGupta. The matching, birthday and the strong birthday problem: a contemporary review. *Journal of Statistical Planning and Inference*, 130(1-2):377–389, 2005.

M. DeGiorgio and J. H. Degnan. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. *Systematic Biology*, 63(1):66–82, 2014.

J. H. Degnan. Modeling hybridization under the network multispecies coalescent. *Systematic Biology*, 67(5):786–799, 2018.

J. H. Degnan and N. A. Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS genetics*, 2(5):e68, 2006.

J. H. Degnan and N. A. Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6):332–340, 2009.

J. H. Degnan and L. A. Salter. Gene tree distributions under the coalescent process. *Evolution*, 59:24–37, 2005.

X. Didelot, R. G. Everitt, A. M. Johansen, and D. J. Lawson. Likelihood-free estimation of model evidence. *Bayesian Analysis*, 6(1):49–76, 2011.

F. Disanto and N. A. Rosenberg. Asymptotic properties of the number of matching coalescent histories for caterpillar-like families of species trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(5):913–925, 2015.

H. H. Fan and L. S. Kubatko. Estimating species trees using approximate bayesian computation. *Molecular Phylogenetics and Evolution*, 59(2):354–363, 2011.

W. Feller. An introduction to probability theory and its applications. *1957*.

W. Feller. *An introduction to probability theory and its applications, vol 2*. John Wiley & Sons, 2008.

REFERENCES

J. Felsenstein. *PHYLIP (phylogeny inference package), version 3.5c.* 1993.

J. Felsenstein. *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA, 2004.

R. A. Fieher. The genetical theory of natural selection, 1930. In *Proc. Roy. Soc. Edinburgh*, volume 42, pages 321–41, 1922.

P. Flajolet, D. Gardy, and L. Thimonier. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, 39(3):207–229, 1992.

A. Gnedin and Y. Yakubovich. On the number of collisions in λ -coalescents. *Electronic Journal of Probability*, 12:1547–1567, 2007.

P. H. Harvey and A. Rambaut. Phylogenetic extinction rates and comparative methodology. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1406): 1691–1696, 1998.

P. H. Harvey, R. M. May, and S. Nee. Phylogenies without fossils. *Evolution*, 48(3):523–529, 1994.

I. Holmes and W. J Bruno. Evolutionary hmms: a bayesian approach to multiple alignment. *Bioinformatics*, 17(9):803–820, 2001.

H. Huang and L. L. Knowles. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of rad sequences. *Systematic Biology*, 65 (3):357–365, 2016.

H. Huang, Q. He, L. S. Kubatko, and L. L. Knowles. Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Systematic Biology*, 59(5):573–583, 2010.

J. P. Huelsenbeck and M. Kirkpatrick. Do phylogenetic methods produce trees with biased shapes? *Evolution*, 50(4):1418–1424, 1996.

T. Janzen, S. Höhna, and R. S. Etienne. Approximate bayesian computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nltt. *Methods in Ecology and Evolution*, 6(5):566–575, 2015.

H. W. Kayondo, S. Mwalili, and J. M. Mango. Inferring multi-type birth-death parameters for a structured host population with application to hiv epidemic in africa. *Computational Molecular Bioscience*, 9(4):108–131, 2019.

A. Kim and J. H. Degnan. PRANC: Ml species tree estimation from the ranked gene trees under coalescence. *Bioinformatics*, 2020. doi: https://doi.org/10.1093/bioinformatics/btaa605.

A. Kim and J. H. Degnan. Heuristics for unrooted, unranked, and ranked anomaly zones under birth- death models. *Molecular Phylogenetics and Evolution*, 161:107162, 2021.

A. Kim, N. A. Rosenberg, and J. H. Degnan. Probabilities of unranked and ranked anomaly zones under birth-death models. *Molecular Biology and Evolution*, 37:1480–1494, 2020.

J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3): 235–248, 1982.

M. S. Klamkin and D. J. Newman. Extensions of the birthday surprise. *Journal of Combinatorial Theory*, 3(3):279–282, 1967.

L. S. Kubatko and J. H. Degnan. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1):17–24, 2007.

L. S. Kubatko, B. C. Carstens, and L. L. Knowles. Stem: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25(7):971–973, 2009.

N. Kutsukake and H. Innan. Detecting phenotypic selection by approximate bayesian computation in phylogenetic comparative methods. In *Modern phylogenetic comparative methods and their application in evolutionary biology*, pages 409–424. Springer, 2014.

A. D Leaché and B. Rannala. The accuracy of species tree estimation under simulation: a comparison of methods. *Systematic Biology*, 60(2):126–137, 2011.

B. Levin. A representation for multinomial cumulative distribution functions. *The Annals of Statistics*, pages 1123–1126, 1981.

L. Liu and S. V. Edwards. Phylogenetic inference in the anomaly zone. *Systematic Biology*, 58:452–460, 2009.

REFERENCES

L. Liu, L. Yu, L. Kubatko, D. K. Pearl, and S. V. Edwards. Coalescent methods for estimating phylogenetic trees. *Molecular Biology and Evolution*, 53(1):320–328, 2009.

W. P. Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 1997.

W. P. Maddison and L. L. Knowles. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, 55(1):21–30, 2006.

W. P. Maddison, P. E. Midford, and S. P. Otto. Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, 56(5):701–710, 2007.

S. Magallon and M. J Sanderson. Absolute diversification rates in angiosperm clades. *Evolution*, 55(9):1762–1780, 2001.

C. L Mallows. An inequality involving multinomial probabilities. *Biometrika*, pages 422–424, 1968.

P. Marjoram and S. Tavaré. Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics*, 7(10):759–770, 2006.

S. Mase. Approximations to the birthday problem with unequal occurrence probabilities and their application to the surname problem in japan. *Annals of the Institute of Statistical Mathematics*, 44(3):479–499, 1992.

G. McCulloch and K. Irvine. Breeding of greater and lesser flamingos at sua pan, botswana, 1998–2001. *Ostrich-Journal of African Ornithology*, 75(4):236–242, 2004.

S. Mirarab, R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. Astral: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17): i541–i548, 2014.

E. K. Molloy and T. Warnow. To include or not to include: the impact of gene filtering on species tree estimation methods. *Systematic Biology*, 67(2):285–303, 2018.

A. Mooers, O. Gascuel, T. Stadler, H. Li, and M. Steel. Branch lengths on birth–death trees and the expected loss of phylogenetic diversity. *Systematic Biology*, 61(2):195–203, 2012.

A. O. Mooers and S. B. Heard. Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology*, 72:31–54, 1997.

P. A. P. Moran. Random processes in genetics. In *Mathematical proceedings of the cambridge philosophical society*, volume 54, pages 60–71. Cambridge University Press, 1958.

S. Nee, R. M. May, and P. H Harvey. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1309): 305–311, 1994.

M. Nei. *Molecular evolutionary genetics.* Columbia university press, 1987.

L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh. Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology. Evolution*, 32:268–274, 2014.

T. S. Nunnikhoven. A birthday problem solution for nonuniform birth frequencies. *The American Statistician*, 46(4):270–274, 1992.

P. Pamilo and M. Nei. Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5):568–583, 1988.

E. Paradis. The distribution of branch lengths in phylogenetic trees. *Molecular Phylogenetics and Evolution*, 94:136–145, 2016.

K. V. Parag and O. G. Pybus. Exact bayesian inference for phylogenetic birth-death models. *Bioinformatics*, 34(21):3638–3645, 2018.

D. L. Rabosky, S. C. Donnellan, A. L. Talaba, and I. J. Lovette. Exceptional among-lineage variation in diversification rates during the radiation of australia's most diverse vertebrate clade. *Proceedings of the Royal Society B: Biological Sciences*, 274(1628): 2915–2923, 2007.

A. Rambaut and N. C. Grassly. Seq-gen: An application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, 13:235–238, 1997.

S. V. Rannala, B.and Edwards, A. Leaché, and Z. Yang. The multi-species coalescent model and species tree inference. *Phylogenetics in the Genomic Era*, page 3–3, 2020.

R. E. Ricklefs. Estimating diversification rates from phylogenetic information. *Trends in Ecology & Evolution*, 22(11):601–610, 2007.

D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147, 1981.

S. Roch and M. Steel. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100: 56–62, 2015.

S. Roch and T. Warnow. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Systematic Biology*, 64(4):663–676, 2015.

S. Roch, M. Nute, and T. Warnow. Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. *Systematic Biology*, 68(2):281–297, 2019.

M. S. Rosenberg and S. Kumar. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. *Molecular Biology and Evolution*, 20(4):610–621, 2003.

N. A. Rosenberg. The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology*, 61(2):225–247, 2002.

N. A. Rosenberg. Counting coalescent histories. *Journal of Computational Biology*, 14(3): 360–377, 2007.

REFERENCES

L. Salichos and A. Rokas. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497:327–331, 2013.

C. Solís-Lemus and C. Ané. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS genetics*, 12(3):e1005896, 2016.

T. Stadler. Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, 267(3):396–404, 2010.

T. Stadler. Simulating trees on a fixed number of extant species. *Systematic Biology*, 60: 676–684, 2011a.

T. Stadler. Inferring epidemiological parameters on the basis of allele frequencies. *Genetics*, 188(3):663–672, 2011b.

T. Stadler. Recovering speciation and extinction dynamics based on phylogenies. *Journal of Evolutionary Biology*, 26(6):1203–1219, 2013a.

T. Stadler. How can we improve accuracy of macroevolutionary rate estimates? *Systematic Biology*, 62(2):321–329, 2013b.

T. Stadler. Package 'treesim'. 2019.

T. Stadler and M. Steel. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. *Journal of Theoretical Biology*, 297:33–40, 2012a.

T. Stadler and M. Steel. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. *Journal of Theoretical Biology*, 297:33–40, 2012b.

T. Stadler, D. Kühnert, S. Bonhoeffer, and A. J. Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv). *Proceedings of the National Academy of Sciences*, 110(1):228–233, 2013.

T. Stadler, J. H. Degnan, and N. A. Rosenberg. Does gene tree discordance explain the mismatch between macroevolutionary models and empirical patterns of tree shape and branching times? *Systematic Biology*, 65:628—639, 2016a.

T. Stadler, J. H. Degnan, and N. A. Rosenberg. Does gene tree discordance explain the mismatch between macroevolutionary models and empirical patterns of tree shape and branching times? *Systematic Biology*, 65(4):628–639, 2016b.

M. Sunnåker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. Approximate bayesian computation. *PLoS Computational Biology*, 9(1):e1002803, 2013.

N. Takahata. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics*, 122(4):957–966, 1989.

M. M. Tanaka, A. R. Francis, F. Luciani, and S. A. Sisson. Using approximate bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, 173(3):1511–1520, 2006.

J. Truszkowski, C. Scornavacca, and F. Pardi. Computing the probability of gene trees concordant with the species tree in the multispecies coalescent. *Theoretical Population Biology*, 137:22–31, 2021.

M. Turelli, N. H. Barton, and J. A. Coyne. Theory and speciation. *Trends in Ecology & Evolution*, 16(7):330–343, 2001.

P. Vachaspati and T. Warnow. Astrid: accurate species trees from internode distances. *BMC Genomics*, 16(10):1–13, 2015.

K. R. Veeramah, A. E. Woerner, L. Johnstone, I. Gut, M. Gut, T. Marques-Bonet, L. Carbone, J. D. Wall, and M. F. Hammer. Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate bayesian computation approach. *Genetics*, 200(1):295–308, 2015.

M. V. Wilson. Is there a characteristic rate of radiation for the insects? *Paleobiology*, 9 (1):79–85, 1983.

S. Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.

Z. Xi, L. Liu, and C. C. Davis. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. *Molecular Phylogenetics and Evolution*, 92:63–71, 2015.

B. Xu and Z. Yang. Challenges in species tree estimation under the multispecies coalescent model. *Genetics.*, 204:1353–1368, 2016.

S. Zhu, J. H. Degnan, S. J. Goldstien, and B. Eldon. Hybrid-lambda: simulation of multiple merger and kingman gene genealogies in species networks and species trees. *BMC Bioinformatics*, 16:1–7, 2015.