Topics in Phylogenetic Species Tree Inference under the Coalescent Model

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Yuan Tian, M.S.

Graduate Program in Evolution, Ecology and Organismal Biology

The Ohio State University

2016

Dissertation Committee:

Dr. Laura Kubatko, Advisor

Dr. Andrea Wolfe

Dr. Bryan Carsens

Dr. Zakee Sabree
© Copyright by

Yuan Tian

2016
Abstract

Phylogenetic tree inference is a fundamental tool to estimate the ancestor-descendant relationships for different species. Currently, it is of great interest to explore the evolutionary relationships for a set of species for which DNA data have been collected and thus accurate and efficient methods are required to estimate phylogenetic trees. However, because the evolutionary relationships can be analyzed at two distinct levels (gene trees and species trees), and it is not necessary for the gene trees and species trees to agree with one another, phylogenetic inference has become increasingly complicated. Incomplete lineage sorting (ILS) is considered to be one of the major factors that cause disagreement between species trees and gene trees. The coalescent process is a widely-accepted model for ILS, and numerous genealogy-based phylogenetic inference methods have been established based on the coalescent model.

In this thesis, coalescent-based methods for phylogenetic tree inference are studied. In Chapter 2, the expected amount of incongruence between gene trees under the same species tree is considered. More specifically, the extent of gene tree incongruence arising from incomplete lineage sorting, as modeled by the coalescent process, is computed. The results in Chapter 2 highlight the fact that substantial discordance among gene trees may occur, even when the number of species is very small. In Chapter 3, a coalescent model for three species that allows gene flow between both pairs of sister populations is proposed, and the resulting gene tree history distribution
is derived. The results suggest conditions under which the species tree and associated parameters, such as the ancestral effective population sizes and the rates of gene flow, are not identifiable from the gene tree topology distribution when gene flow is present, but indicate that the coalescent history distribution may identify the species tree and associated parameters. In Chapter 4, a rooting method based on the site pattern probabilities under the coalescent model is developed. The proposed technique provides a method to root every four-taxon species tree within a larger species tree of more than four taxa. The inferred roots for the four-taxon subtrees are then used together to estimate the root for the larger species tree. This rooting method is a computationally feasible method, and is the first method proposed to root a species tree that explicitly incorporates the coalescent process.
This thesis is dedicated to my husband and parents, for their constant love and support during the pursuit of my graduate degree.
Acknowledgments

I wish foremost to thank my dear advisor Dr. Laura Kubatko for her guidance, enthusiasm, and support throughout my graduate study. Without her support and encouragement, I would not be able to finish a dual-degree program and enjoy the fun of research. She is one of the most important people in my life and I view her as an intelligent advisor, a nice friend, and a great role model. During my graduate study, Laura has always been helpful, understanding and supportive. No words can express how fortunate I am to have her as my PhD advisor.

I would also like to thank the members of my dissertation and Ph.D candidacy exam committee, Dr. Andrea Wolfe, Dr. Bryan Carsens, and Dr. Zakee Sabree, for their support, advice, and helpful suggestions throughout the completion of my doctoral work.

I appreciate the great help from Dr. Matthew Hahn and Dr. James Pease for their willingness to share their mosquito data, and Dr. Pease for assistance in providing the data. The mosquito data set provides me a good test for my dissertation work. Besides, I would like to thank Dr. David Swofford for providing delightful discussion to my research projects.

In addition, I would like to thank the co-workers and colleagues from Department of Evolution, Ecology and Organismal Biology at The Ohio State University, including
but not limited to Paul Blischak, Aaron Wenzel, Jeffery Gaither, Alejandro Otero Bravo, Rosa Rodriguez Pena, Benjamin Stone, Emiko Waight, and Coleen Thompson.

I also want to acknowledge the following grant that supported my research: National Science Foundation (NSF), Division Of Environmental Biology, under DEB-1455399.

Finally, I would like to acknowledge the supports of my beloved husband and parents. None of this thesis work would be possible without their constant support.
Vita

2009 .................................................. B.S. Biological Sciences, Tsinghua University

2012 .................................................. M.S. Plant Cellular and Molecular Biology, The Ohio State University

2015 .................................................. M.S. Statistics, The Ohio State University

2013-present ................................. Graduate Teaching/Research Associate, The Ohio State University.

Publications


Fields of Study

Major Field: Evolution, Phylogenetics, Statistics
# Table of Contents

Abstract .......................................................................................................................... ii

Dedication ........................................................................................................................ iv

Acknowledgments ......................................................................................................... v

Vita .................................................................................................................................. vii

List of Tables .................................................................................................................. xii

List of Figures ............................................................................................................... xiii

1. Introduction .............................................................................................................. 1
   1.1 Gene trees and nucleotide substitution models ................................................. 2
   1.2 The coalescent process and computing site pattern probabilities ...................... 7
   1.3 Overview of the dissertation .............................................................................. 12
      1.3.1 Phylogenetic tree incongruence ................................................................. 12
      1.3.2 Hybridization and gene flow ..................................................................... 14
      1.3.3 Algebraic statistical method in species tree inference ................................. 15

2. Expected pairwise congruence among gene trees under the coalescent model 17
   2.1 Introduction ......................................................................................................... 18
   2.2 Methods ............................................................................................................... 21
      2.2.1 Gene tree incongruence probability for species trees with varying internal
            branch lengths ............................................................................................... 22
      2.2.2 Distribution of pairwise RF distance among gene trees ............................ 23
   2.3 Results .................................................................................................................. 24
      2.3.1 Gene tree incongruence increases with an increasing number
            of taxa ........................................................................................................... 24
2.3.2 Gene tree incongruence decreases with increasing branch lengths of the species tree .............................................. 25
2.3.3 Distribution of RF distance between two gene trees .......... 27

2.4 Discussion .................................................................. 30

3. Distribution of gene tree histories under the coalescent model with gene flow 35

3.1 Introduction ............................................................... 36
3.2 Materials and Methods .................................................. 41
  3.2.1 A model for three species with gene flow ...................... 41
  3.2.2 Maximum likelihood parameter estimation .................... 46
3.3 Theory ....................................................................... 47
  3.3.1 The probability distribution of coalescent histories ........ 47
  3.3.2 Implementation details and parameter scaling ............... 54
3.4 Results ...................................................................... 56
  3.4.1 Gene flow between sister species produces different distributions of gene tree histories .................................. 56
  3.4.2 Variation in the gene tree history distribution as a function of the rate of gene flow ........................................ 60
  3.4.3 Different distributions of gene tree histories may share an identical gene tree topology distribution .................. 66
  3.4.4 Simulation studies for maximum likelihood parameter estimation ................................................................. 69
  3.4.5 Application to an empirical Afrotropical mosquito data set ................................................................. 74
3.5 Discussion ................................................................. 77
  3.5.1 Identifying the species tree from the gene tree topology distribution in the presence of gene flow ..................... 77
  3.5.2 Applying the method to empirical data ......................... 79
  3.5.3 Future directions ...................................................... 81
  3.5.4 Conclusions .......................................................... 83

4. Rooting phylogenetic trees under the coalescent model using site pattern probabilities ............................................. 102

4.1 Introduction ............................................................... 103
4.2 Methods ................................................................. 106
  4.2.1 Method for Rooting Phylogenetic Trees by Site Pattern Probabilities ................................................................. 106
  4.2.2 Formal hypothesis tests .............................................. 108
  4.2.3 Simulation Studies ................................................... 111
  4.2.4 Application to Larger Species Trees ............................ 113
4.3 Results ...................................................................... 115
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.1</td>
<td>Accuracy of the method for rooting phylogenetic trees</td>
<td>115</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Application to an eight-taxa North American rattlesnake data set</td>
<td>119</td>
</tr>
<tr>
<td>4.4</td>
<td>Discussion</td>
<td>120</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Rooting phylogenetic trees under different nucleotide substitution models</td>
<td>122</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Rooting phylogenetic trees using multi-loci DNA sequence data</td>
<td>124</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Control of familywise error rate</td>
<td>124</td>
</tr>
</tbody>
</table>

5. Discussion and Future Work ........................................ 127

Bibliography ............................................................. 132

Appendices

A. Derivation of the gene tree history distribution under the coalescent model 157
List of Tables

Table 1  Relationships expected among site pattern probabilities for various root positions. ........................................ 107

Table 2  Test results and conclusions for rooting a four-taxon phylogenetic tree. ...................................................... 111

Table 3  Rooting results for the North American rattlesnake data set .......................... 122
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>An example of one possible phylogenetic relationship between the species tree (outline tree) and the gene tree (embedded tree) for a four-leaf tree</td>
<td>3</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Sites, SNPs, and coalescent independent sites</td>
<td>4</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Gene tree congruence for species trees with different numbers of taxa.</td>
<td>26</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Gene tree congruence for four-taxon and five-taxon species trees</td>
<td>28</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Distribution of pairwise RF distances among gene trees for the five-taxon species tree: $(((A : T_0, B : T_0) : T_1, C : (T_0 + T_1)) : T_2, D : (T_0 + T_1 + T_2)) : T_3, E : (T_0 + T_1 + T_2 + T_3))$</td>
<td>29</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Probability distribution of pairwise RF distances between gene trees in a seven-taxon species tree.</td>
<td>31</td>
</tr>
<tr>
<td>Figure 7</td>
<td>The model species tree $((A, B), C)$ for three species with gene flow between sister species.</td>
<td>42</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Possible gene tree histories with three different gene tree topologies</td>
<td>43</td>
</tr>
<tr>
<td>Figure 9</td>
<td>The probability distribution of the gene tree histories under species trees with scaled speciation times $T_1 = 1.6, T_2 = 2.4$.</td>
<td>57</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Probability distributions of the gene tree histories for the model of three species with gene flow between sister species.</td>
<td>62</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Probability distributions of the gene tree histories for the model of three species with gene flow between sister species.</td>
<td>64</td>
</tr>
</tbody>
</table>
Figure 12  Coalescent history distributions for eight species trees with dif-
ferent coalescent rates and gene flow rates. .......................... 68

Figure 13  Contour plot for one simulated data set showing the log likeli-
hood as a function of the scaled coalescent rate C and the gene flow
rate M. ................................................................. 70

Figure 14  Maximum likelihood estimates of the scaled coalescent rates
and the gene flow rates obtained from the simulated data sets under
the three species coalescent model with gene flow between both sister
taxa. ................................................................. 71

Figure 15  Contour plots of the log likelihood for the coalescent rate $2/\theta$
and the gene flow rate $M_2$ at four different levels of gene flow rate $M_1$. 76

Figure 16  The probability distribution of the gene tree histories under
species trees with different scaled speciation times. ................. 84

Figure 17  Probability distributions of the gene tree histories for the model
of three species with gene flow between sister species (equal population
size, short speciation times, constant $M_2$). ......................... 86

Figure 18  Probability distributions of the gene tree histories for the model
of three species with gene flow between sister species (equal population
size, long speciation times, constant $M_2$). .......................... 88

Figure 19  Probability distributions of the gene tree histories for the model
of three species with gene flow between sister species (unequal popu-
lation size, short speciation times, constant $M_2$). ....................... 90

Figure 20  Probability distributions of the gene tree histories for the model
of three species with gene flow between sister species (unequal popu-
lation size, long speciation times, constant $M_2$). ....................... 92

Figure 21  Probability distributions of the gene tree histories for the model
of three species with gene flow between sister species (equal population
size, short speciation times, constant $M_1$). .......................... 94

Figure 22  Probability distributions of the gene tree histories for the model
of three species with gene flow between sister species (equal population
size, long speciation times, constant $M_1$). .......................... 96
Figure 23  Probability distributions of the gene tree histories for the model of three species with gene flow between sister species (unequal population size, short speciation times, constant $M_1$)  

Figure 24  Probability distributions of the gene tree histories for the model of three species with gene flow between sister species (unequal population size, long speciation times, constant $M_1$)  

Figure 25  Example species trees  

Figure 26  Example of scoring potential root positions on larger phylogenies  

Figure 27  Accuracy of the rooting method  

Figure 28  Application of the rooting method to the rattlesnake data set
Chapter 1: Introduction

Phylogenetic trees are graphs that display the evolutionary relationships among a collection of biological entities, such as species or genes (Bryant and Steel, 1995; Nuu-tila and Soisalon-Soininen, 1994; Page, 2001; Felsenstein, 1981; Sattath and Tversky, 1977). A phylogenetic tree includes tips (leaves), internal nodes, and branches (Figure 1). The leaves represent present-day species, populations or genes, and the internal nodes represent their common ancestor. The branches connect nodes to nodes and nodes to leaves. Sometimes, the branch lengths may be interpreted to correspond to evolutionary time (De Queiroz, 2013). A phylogenetic tree that does not include the information of branch lengths is sometimes called a tree topology. In evolutionary biology, phylogenetic tree inference is a fundamental tool to estimate the ancestor-descendant relationships for different species (Felsenstein, 2004; Swofford, 2001).

Currently, it is of great interest to explore the evolutionary relationships for a set of species for which DNA data have been collected. Besides representing the species relationships, phylogenetic inference can also identify the histories of populations (Edwards, 2009), the epidemiological dynamics of pathogens (Marra et al., 2003; Grenfell et al., 2004; Bobkov et al., 1994; Li et al., 2000), and events on the phylogeny such as horizontal gene transfer (Andersson et al., 2003; Ochman et al., 2000). Moreover, phylogenetic analyses are useful in forensic medicine (Hillis and Huelsenbeck, 1994;
Ou et al., 1992), comparative (e.g., ecological or physiological) studies (Givnish, 1987; Felsenstein, 1985), and in conservation biology (May-Collado and Agnarsson, 2011). Recently, the amount of available DNA sequence data has increased dramatically due to advances in sequencing technology. Thus, more accurate and efficient methods are required to estimate phylogenetic trees in order to study the evolutionary patterns and processes from the information contained in representative DNA sequences (Yang and Rannala, 2012).

1.1 Gene trees and nucleotide substitution models

Evolutionary relationships can be analyzed at two distinct levels. One is the level of the species (or populations), and the other is the level of the individual genes. A gene tree is constructed from DNA sequences for a genetic locus, while a species tree represents the actual evolutionary pathway of the species involved (Pamilo and Nei, 1988). It is clear that phylogenetic inference on the species level and the gene level cannot be considered separately, since the evolutionary histories for the individual genes (gene trees) are constrained by the histories of the species (species trees) (Maddison, 1997; Edwards et al., 2007). Note that it is not necessary for the gene trees and species trees to agree with one another (Figure 1) (Pamilo and Nei, 1988; Nichols, 2001). Reasons for this disagreement between species trees and gene trees include horizontal gene transfer, hybridization, undetected paralogy, and incomplete lineage sorting (deep coalescence) (Maddison, 1997; Meng and Kubatko, 2009).

To explore the evolutionary relationships among DNA sequences, it is widely accepted that DNA sequences evolve through a nucleotide substitution processes over
Figure 1: An example of one possible phylogenetic relationship between the species tree (outline tree) and the gene tree (embedded tree) for a four-leaf tree. In this figure, speciation times are denoted as $\tau$’s and $\theta$’s represent the effective population sizes.
Figure 2: (a) 15 contiguous sites in the aligned DNA sequence of the COS4270 gene for five of the *Penstemon* species provided by Wolfe et al. (Wolfe *et al.*, 1998a,b); (b) Five coalescent independent sites selected from five distinct genes for five of the *Penstemon* species provided by Wolfe et al. (Wolfe *et al.*, 1998a,b).

Modeling the nucleotide substitution processes allows comparison of two or more DNA sequences that evolved from a common ancestral DNA sequence. To compare multiple DNA sequences, a widely used method is to make a DNA sequence alignment, which is a way of arranging the sequences of DNA to match regions of similarity between the sequences. For instance, an aligned DNA sequence data set is shown in Figure 2a. In this data set, each column is defined as a *site*, and corresponds to an evolutionary process whereby these nucleotides evolved from a single ancestral nucleotide (not shown in the figure). A site is considered to be a *single nucleotide polymorphism* (*SNP*) when the nucleotides in a site are not identical (highlighted in Figure 2a). When a set of sites under the coalescent model are assumed to freely recombine, they are defined as *coalescent independent sites* (Figure 2b).

Multiple nucleotide substitution models have been proposed to model the process of nucleotide substitution using continuous-time Markov chains, which assume a nucleotide substitution rate matrix that specifies the rate of change from one nucleotide to another under the assumption that the future state of the Markov chain depends...
only on the current state of the model. The Jukes-Cantor (JC69) model (Jukes and Cantor, 1969) is the simplest of the commonly used models, assuming equal base frequencies and equal mutation rates for all bases. Kimura (1980) introduced a model, called K80, that distinguishes transition (substitutions between the two purines, \(A\) and \(G\), or between the two pyrimidines, \(C\) and \(T\)) and transversion (substitutions between a purine and a pyrimidine) rates. The K80 model also assumes equal base frequencies. In 1981, Felsenstein proposed the F81 model (Felsenstein, 1981), which is an extension of the JC69 model. In the F81 model, base frequencies are allowed to vary from 0.25. The K80 and F81 model were then unified by Hasegawa, Kishino and Yano in 1985 as the HKY85 model (Hasegawa et al., 1985), in which the rate of transitions and transversions are different, and base frequencies are allowed to be unequal. Several models that are similar to HKY85 have been developed, such as T92 (Tamura, 1992) and TN93 (Tamura and Nei, 1993). Another popular nucleotide substitution model is the Generalized Time-Reversible (GTR) model (Tavare, 1986). The GTR model is the most general model, in which the base frequencies are unequal, and different rates are allowed for the six possible types of substitutions among the bases (Posada, 2003).

To further explore the nucleotide substitution process under the coalescent model, consider the simplest JC 69 model. The JC69 model assumes equal base frequencies with \(\pi_k = \frac{1}{4}\) for \(k \in \{A, C, G, T\}\) and equal mutation rates \(\mu\). Thus, with a Markov model of evolution along time (Jukes and Cantor, 1969), the instantaneous rate matrix for the JC69 model is:
\[
Q = \begin{pmatrix}
A & G & C & T \\
\frac{-3}{4} \mu & \frac{1}{4} \mu & \frac{1}{4} \mu & \frac{1}{4} \mu \\
\frac{1}{4} \mu & \frac{-3}{4} \mu & \frac{1}{4} \mu & \frac{1}{4} \mu \\
\frac{1}{4} \mu & \frac{1}{4} \mu & \frac{-3}{4} \mu & \frac{1}{4} \mu \\
\frac{1}{4} \mu & \frac{1}{4} \mu & \frac{1}{4} \mu & \frac{-3}{4} \mu
\end{pmatrix}
\]

From the matrix one can compute the probability of a nucleotide changing from initial state \( k \) to final state \( l \) as a function of the branch length \( t \) using standard Markov chain theory (Meyn and Tweedie, 2012) to get:

\[
p_{kl}(\nu) = \begin{cases}
\frac{1}{4} + \frac{3}{4} e^{-4 \nu/3} & \text{if } k = l \\
\frac{1}{4} - \frac{1}{4} e^{-4 \nu/3} & \text{if } k \neq l
\end{cases}
\]

(1.1)

Once the nucleotide substitution model is specified, the probability of any combination of the nucleotides within a site along a fixed time can be calculated as the site pattern probability. For instance, the 4-taxon gene tree \( g = (G, t) \) as shown in Figure 1 with 6 branches and particular observation \( k_A \ldots k_D \), \( k_M \) represents the nucleotide state of species \( M \), and \( k_M \in \{A, C, G, T\} \) at one site, the site pattern probability \( p_{k_A k_B k_C k_D} \) is

\[
p_{k_A k_B k_C k_D} = \sum_{S_3} \sum_{S_2} \sum_{S_1} \frac{1}{4} \left( p_{k_{S_2}, S_3}(t_5) \right) \left( p_{k_{S_3}, D}(t_6) \right) \left( p_{k_{S_1}, S_2}(t_4) \right) \left( p_{k_{S_2}, A}(t_1) \right) \left( p_{k_{S_1}, B}(t_2) \right) \left( p_{k_{S_1}, C}(t_3) \right)
\]

(1.2)

where \( S_1, S_2, \) and \( S_3 \) represent the ancestor species, and \( t_m \ (m = 1, 2, \ldots, 6) \) represent the branches as shown in Figure 1.

With the site pattern probability calculated under an assumed gene tree, it is straightforward to compute the likelihood of the gene tree from the multinomial probability of the observed counts of site pattern frequencies (Chifman and Kubatko, 2014, 2015). As a result, the gene trees can be estimated under a maximum likelihood (ML) framework (Felsenstein, 1981, 2004).
1.2 The coalescent process and computing site pattern probabilities

Incomplete lineage sorting (ILS) is considered to be one of the major factors that can cause disagreement between the species trees and gene trees (Pamilo and Nei, 1988; Degnan and Rosenberg, 2006). It is also thought to be a critical issue that decreases the accuracy in estimating species trees with large multilocus data sets (Pamilo and Nei, 1988). For a species tree with more than four taxa, when the gene trees are estimated independently for each locus, it is possible that the gene tree with the highest frequency disagrees with the species tree, thereby identifying an incorrect species tree if a democratic vote procedure is used to infer the species tree from the observed gene tree (Degnan and Rosenberg, 2006). One way to analyze large multilocus data sets is to concatenate all of the DNA sequences as a whole to estimate a single gene tree (Hackett et al., 2008; Rokas et al., 2003). However, this reconstruction procedure was shown to be biased statistically (Kubatko and Degnan, 2007; Roch and Steel, 2014).

An alternative way to study ILS is to model it with the coalescent process (Edwards, 2009; Knowles and Kubatko, 2010), which is a retrospective model of population genetics. The coalescent model is based on tracing the evolutionary history of sampled genes by considering the time from the present back to their most recent common ancestor (Kingman, 2000), and can be derived as the limiting distribution (as the population size becomes large) that results from the Wright-Fisher and other commonly-used population genetics models (Wakeley, 2009; Kingman, 1982a; Takahata and Nei, 1985) Under the coalescent model, the probability distribution of gene
trees given a fixed species tree topology and branch lengths can be computed (Degen
nan and Salter, 2005). The coalescent model is also used as the basis for different
methods to estimate species trees using either multi-locus DNA sequence data or a
set of observed gene trees (Kubatko et al., 2009; Than et al., 2007; Liu and Pearl,
2007; Liu et al., 2010; Heled and Drummond, 2010). These methods have been widely
applied to the multi-locus data sets that are commonly produced by next generation
sequencing techniques. Notably, under the coalescent model, the expected time be-
tween two coalescence events (times at which two alleles arise from a single ancestor)
increases exponentially back in time (Kingman, 1982a,b; Tavaré, 1984).

The coalescent process has been so widely applied in part because it is believed
that ILS is a predominant cause of the incongruence observed between gene trees
and species trees (Liu et al., 2010). Indeed, the predictions made by the coalescent
model in terms of the distribution of gene trees are consistent with several observed
data sets (Ebersberger et al., 2007; Ane, 2010; Kubatko et al., 2011). Based on
coalessent theory, numerous genealogy-based phylogenetic inference methods have
been developed (Edwards, 2009; Rannala and Yang, 2008; Knowles and Kubatko,
2010). A set of branch-length-based methods (Maximum Tree/GLASS (Liu et al.,
2010; Mossel and Roch, 2010), STEAC (Liu et al., 2009), and STEM (Kubatko,
2009)) estimate the species evolutionary relationships by assuming that each locus
has one single underlying gene tree for the multi-locus data, and using estimates of
these gene trees to obtain an overall species tree estimate. Similarly, another group
of methods also use the estimated gene trees for each locus (by maximum likelihood
(ML) (Felsenstein, 1981) or neighbor joining (Saitou and Nei, 1987)), but only the
gene tree topologies are utilized. These methods include ASTRAL (Mirarab et al.,
2014; Mirarab and Warnow, 2015), Minimization of Deep Coalescences (Maddison and Knowles, 2006; Than and Nakhleh, 2009), MP-EST (Liu et al., 2010), ST-ABC (Fan and Kubatko, 2011), STAR (Liu et al., 2009), STELLS (Wu, 2012), and NJst (Liu and Yu, 2011). These methods are efficient in estimating species trees, but their accuracy mostly depends on the accurate estimation of gene trees. Furthermore, a group of Bayesian approaches (*BEAST (Heled and Drummond, 2010), BEST (Liu and Pearl, 2007), and SNAPP (Bryant et al., 2012)) estimate both species trees and gene trees simultaneously by completely using the data and corresponding gene tree prior distributions. Generally, the Bayesian approaches provide a comprehensive estimation of gene tree topology, population sizes, and parameters of substitution models, as well as the species trees, but the huge computational cost that results from the use of MCMC to estimate the posterior distribution limits their application.

The gene trees nested in a species tree (Figure 1) can be modeled by the coalescent process (Kingman, 1982b; Tajima, 1983; Takahata and Nei, 1985; Pamilo and Nei, 1988; Wakeley, 2009). More specifically, the coalescent process models a set of coalescent events for a sample of lineages back in time with an exponential distribution. For example, the probability density of the time to coalescence of two lineages among $j$ lineages is

$$f(t_j) = \frac{j(j - 1)}{2} \frac{2}{\theta} e^{\frac{-j(j - 1)}{2\theta} t_j}, t_j > 0.$$  \hspace{1cm} (1.3)

Here the parameter $\theta$ represents the effective population size and is given by $\theta = 4N\mu$, with the effective population size denoted as $N$ and the mutation rate denoted as $\mu$.

The gene tree probability density can thus be computed for any given species tree (Rannala and Yang, 2003). Denote an $n$-taxon species tree $S$ with the speciation
times denoted as $\tau = (\tau_1, \ldots, \tau_{n-1})$. Consider a population $i$ within the species tree, and define the time at the beginning of the branch (closest to the present time) as $\tau_i$, and the time at the end (closest to the root) as $\tau_{pi}$. Thus, the length of branch $i$ is $\tau_{pi} - \tau_i$. Suppose there are $u_i$ lineages entering the population at time $\tau_i$, and $v_i$ lineages leaving the population at time $\tau_{pi}$, then the number of coalescent events within branch $i$ is $w_i = u_i - v_i$. Ordering these coalescent events back from the present, and letting the vector $t^i = (t^i_1, t^i_2, \ldots, t^i_{w_i})$ denote the times of the $w_i$ coalescent events, the density of $t^i$ can be written as

$$f(t^i|\theta_i, \tau_{pi}, \tau_i) = \exp\left\{-\frac{2}{\theta_i} \left(\frac{v_i}{2}\right)(\tau_{pi} - t^i_{w_i})\right\} \prod_{w=1}^{w_i} \frac{2}{\theta_i} \exp\left\{-\frac{2}{\theta_i} \left(\frac{u_i - w + 1}{2}\right)(t^i_w - t^i_{w-1})\right\} \quad (1.4)$$

where $t^i_0 = \tau_i$ (Rannala and Yang, 2003).

Since the coalescent process within a population is described in Equation (1.4), we can derive the conditional density of the whole gene tree $(G, t)$ given the species tree $(S, \tau)$. Note that $G$ is the gene tree topology and $t$ represents the coalescent times of the gene tree. Given that the numbers of $u_i$ and $v_i$ are specified, the coalescent processes for each population are independent (Rannala and Yang, 2003). Consequently, the density of the entire gene tree $g = (G, t)$ given a species tree $(S, \tau, \Theta)$ can be obtained by multiplying the densities of all individual populations:

$$f(g|S, \tau, \Theta) = \prod_{i=1}^{n-2} f(t^i|\theta_i, \tau_{pi}, \tau_i). \quad (1.5)$$

Here $\Theta$ is a vector representing the population size parameters for each branch.

With the site pattern probabilities for a gene tree derived in Equation (1.2), and the gene tree density under a given species tree as in Equation (1.5), the species tree
site pattern probabilities can be computed by weighting the gene tree site pattern probabilities based on the probability of each gene tree under the coalescent model with a given species tree, and then summing over all possible gene trees (Chifman and Kubatko, 2014, 2015). Though the site pattern probabilities can be computed for an arbitrary number of species, here we are only interested in the four-taxon (quartet) case since quartets are used as minimal units to estimate the species tree in our work.

For any four-taxon species tree (one example is shown in Figure 1), the nucleotide observation at the tips of the phylogeny is denoted by $\sigma(i_1, i_2, i_3, i_4)$. Note that $\sigma$ is a permutation of $i_1, i_2, i_3, i_4$ representing a re-ordering of nucleotides. Thus, the probability of any observation $i_1, i_2, i_3, i_4$ given the species tree $(S, \tau, \Theta)$ is

$$p^*_{i_1i_2i_3i_4}(S, \tau, \Theta) = \sum_G \int_t p_{\sigma(i_1, i_2, i_3, i_4)}(G, t) \cdot f(G, t|S, \tau, \Theta) dt,$$

(1.6) Note that the species tree site pattern probabilities are obtained by taking the sum over all gene trees that are integrated over the branch lengths. Under a given species tree $(S, \tau, \Theta)$, the probability distribution of all possible site patterns can be computed for different nucleotide substitution models as described in Equation (1.2). Specifically, for the simplest JC69 model, the identical base frequencies and the constant nucleotide substitution rate produce identical site pattern probabilities in many cases. For instance, given a four-taxon tree, the total number of possible site patterns is $4^4 = 256$, but there are only 15 distinct site pattern probabilities (Chifman and Kubatko, 2014, 2015).
1.3 Overview of the dissertation

1.3.1 Phylogenetic tree incongruence

Under the coalescent model, it is possible to observe different gene trees within the same species tree, and the gene tree topology with the highest probability need not match the species tree. Thus, the ability to compute gene tree probability distributions provides several important insights into the problem of multi-locus species tree estimation. Degnan and Salter (2005) derived a method for computing the distribution of gene tree topologies given a bifurcating species tree with an arbitrary number of taxa, when one gene is sampled for each species. This approach used the concept of gene tree coalescent histories, which are gene tree topologies together with an assignment of coalescent events on the gene tree topology to specific intervals of the species tree. This method was implemented in the computer program COAL. When a given species tree (with branch lengths in coalescent units) is input into COAL, the probability of each possible gene tree topology will be calculated.

Although it is widely appreciated that gene trees may differ from the overall species tree, few papers highlight that incongruence exists among gene trees under the same species tree. Moreover, the extent of this incongruence is rarely quantified and discussed. Since the incongruence between gene trees and species trees, or among different gene trees, can be either explained by ILS, or may be due simply to random chance or sampling error, it is of great interest to measure the extent of phylogenetic tree incongruence and compare it to the expected amount of phylogenetic tree incongruence caused by ILS.

Qualitative comparison of phylogenetic trees can be made directly by visualizing the differences between two trees with tanglegrams and exploring the differences in
collections of trees. Examples of approaches are consensus networks (Holland et al., 2005) for unrooted trees, and DensiTree plots (Bouckaert, 2010) for rooted trees. However, such visualization methods become unwieldy and uninformative when dealing with larger trees or multiple trees (Kendall and Colijn, 2015). Alternatively, metric-based quantitative tree comparison methods can be used to measure the extent of phylogenetic tree incongruence. The Robinson-Foulds (RF) unweighted metric (Robinson and Foulds, 1981), also known as the “symmetric difference, or “RF distance”, is one of the most widely used metric-based quantitative tree comparison measures (Kendall and Colijn, 2015). The RF distance measures the distance between two phylogenetic trees, and is computed as the sum of the number of partitions of data implied by the first tree but not the second tree and the number of partitions of data implied by the second tree but not the first tree (Robinson and Foulds, 1981). A larger RF distance indicates more incongruence between two phylogenetic trees.

In Chapter 2, we consider the expected amount of gene tree incongruence arising from ILS, and compute the probability that two gene trees randomly sampled from the same species tree agree with one another. Furthermore, we compute the distribution of the RF distance between two randomly sampled gene trees from the same species tree with three to eight taxa. The study highlights the fact that substantial discordance may occur, even when the number of species is very small, which has implications both for larger taxon samples and for any method that uses estimated gene trees as the basis for further statistical inference.
1.3.2 Hybridization and gene flow

Currently, the gene tree probability distribution can be easily computed by COAL (Degnan and Salter, 2005) for any fixed species tree. However, the method implemented in COAL assumes no hybridization or gene flow following speciation. Hybridization describes mating between two individuals within different populations that may generate a hybrid individual with mixed genetic information from both populations. New species may be created if hybridization lasts for many generations between two populations, thereby generating fertile offspring. On the other hand, gene flow may exist among species without creating new hybrid species. Actually, hybridization and gene flow exists widely in a variety of species and is now considered to be a significant mechanism in generating new species (Rieseberg, 1997; Gross and Rieseberg, 2005; Buerkle et al., 2000; Bullini, 1994; Nolte et al., 2005; DeMarais et al., 1992; Gompert et al., 2006; Schwarz et al., 2005; Mavarez, 2006; Meyer et al., 2006; Mallet, 2007; Seehausen, 2004; Mallet, 2005, 2007; Baack and Rieseberg, 2007).

Thus, it is important to incorporate the possibility of hybridization and gene flow into methodology for phylogenetic inference. However, a normal bifurcating phylogenetic tree (Figure 1) cannot represent hybridizing species or species with gene flow, since the hybrid species or species with gene flow contain genetic information from two distinct species. Thus, it is important to establish statistical models that incorporate both hybridization/gene flow and ILS jointly.

In Chapter 3, we propose a coalescent model for three species that allows gene flow between both pairs of sister populations. The model is developed based on an Isolation-with-migration (IM) model (Hey and Nielsen, 2004; Hey, 2010) to model both population splitting and gene flow. The model is formulated using a Markov
chain representation, which allows use of matrix exponentiation to compute analytical expressions for the probability density of gene tree genealogies. This method is applied to an Afrotropical mosquito data set (provided by Fontaine et al., 2015) to demonstrate an application of our method to the analysis of empirical data.

1.3.3 Algebraic statistical method in species tree inference

In most species tree inference approaches, gene trees are estimated first and are assumed known in the following analysis. However, it is highly possible that the estimated gene trees are biased or not fully informative, because they are often based on short sequences with few variable sites (Brower et al., 1996). As a result, gene tree estimation errors may potentially become a severe issue in species tree inference. Recently, an algebraic statistical method was developed for species tree inference without the pre-requisite of estimating gene trees. This method, SVDQuartets, is a full-data method based on the site pattern probabilities under a species tree (Chifman and Kubatko, 2014, 2015). The method provides efficient species tree inference because it does not require the prior estimation of gene trees that may potentially introduce estimation errors. More specifically, SVDQuartets estimates the site pattern probabilities for each sampled quartet, and then an SVD score is calculated for a matrix of estimated site pattern probabilities corresponding to each of the three possible splits (AB\vert CD, AC\vert BD, or AD\vert BC) in a quartet. The true split is inferred to be the one with the lowest score based on the genome-scale data. The inferred splits for a set of sampled quartets can be input to a quartet amalgamation algorithm to construct a complete species tree for more than four species. Note that SVDQuartets assumes
free recombination among the sites. In other words, it is designed for coalescent independent sites (Figure 2b), but previous simulation studies and real-data analyses also indicated good performance of SVDQuartets in analyzing multi-loci DNA sequence data. Notably, SVDQuartets can be applied to estimate a phylogenetic tree under typical nucleotide substitution models (JC69, HKY85, GTR, etc.) (Chifman and Kubatko, 2014, 2015). Also, SVDQuartets is suitable for the case of variable substitution rates across the sites (the substitution rates are drawn from an arbitrary Gamma distribution) and a proportion of invariant sites (Yang, 1993, 1994). The computational cost of SVDQuartets is significantly lower than other methods for species tree inference under the coalescent model. Analyses of empirical data indicate that SVDQuartets can handle data sets of 50 or more taxa in a reasonable time (Kubatko and Chifman, 2015).

The information provided by site pattern probabilities is not limited to estimating unrooted bifurcating phylogenetic trees. In Chapter 4, we develop a new method for rooting quartet species trees under the coalescent model, by developing a series of hypothesis tests for rooting species-level phylogenies using site pattern probabilities. The power of this method is examined by simulation studies and by application to an empirical North American rattlesnake data set. Our study establishes a computationally practical method for rooting species trees. The method will benefit numerous evolutionary studies that require rooting a phylogenetic tree without having to specify outgroups.
Chapter 2: Expected pairwise congruence among gene trees under the coalescent model

Although it is widely appreciated that gene trees may differ from the overall species tree and from one another tree due to various evolutionary processes (e.g., incomplete lineage sorting, horizontal gene transfer, etc.) (Pamilo and Nei, 1988; Takahata, 1989; Hein, 1993; Maddison, 1997), the extent of this incongruence is rarely quantified and discussed. Here we consider the expected amount of incongruence arising from incomplete lineage sorting, as modeled the coalescent process. In particular, we compute the probability that two gene trees randomly sampled from the same species tree agree with one another as well as the distribution of the Robinson-Foulds distance between them, for species trees with three to eight taxa. We demonstrate that, as expected under the coalescent model, the amount of discordance is affected by species tree-specific factors such as the speciation times and the effective population sizes for the species under consideration. Our results highlight the fact substantial discordance may occur, even when the number of species is very small, which has implications both for larger taxon samples and for any method that uses estimated gene trees as a basis for further statistical inference. We conclude that the amount of incongruence is substantial enough that such methods may need to be modified to account for variability in the underlying gene trees.
2.1 Introduction

Incomplete lineage sorting (ILS) has long been recognized to be a predominant cause of substantial variation in the evolutionary trees for individual genes, leading to incongruence among gene trees (Pamilo and Nei, 1988; Takahata, 1989; Hein, 1993; Maddison, 1997; Sang and Zhong, 2001; Kubatko, 2009; Liu et al., 2010; Bayzid and Warnow, 2012). Coalescent theory (Kingman, 1982a,b; Tajima, 1983; Tavaré, 1984; Takahata and Nei, 1985; Pamilo and Nei, 1988; Rosenberg, 2002; Rannala and Yang, 2003; Degnan and Salter, 2005) is commonly used to model ILS, and the predictions concerning agreement among gene trees made by the coalescent model are becoming increasingly well-understood. For example, it is widely appreciated that the extent of gene tree incongruence depends on characteristics of the species tree, such as branch lengths and effective population sizes (Pamilo and Nei, 1988; Rosenberg, 2002; Rannala and Yang, 2003; Degnan and Salter, 2005), suggesting the importance of considering the coalescent process in phylogenetic studies (Degnan and Salter, 2005; Degnan and Rosenberg, 2006). Indeed, most current methods of multi-locus phylogenetic inference incorporate the coalescent process to model ILS (Edwards et al., 2016). In this study, we consider the extent of congruence between a pair of gene trees that is expected under the coalescent process for a small number of taxa. Our study is motivated by four current challenges in empirical multilocus phylogenetics.

First, we consider the problem of assessing the extent to which empirical data fit the predictions made by the coalescent model. In particular, a recent study (Simmons et al., 2016) examined eight empirical data sets and presented the pairwise congruence among gene trees for each of these studies. Simmons et al. (2016) reported that the average pairwise congruence among gene trees varied greatly both between studies and
sometimes within a study. However, the study did not consider the extent of pairwise congruence that would be *expected* among gene trees under the coalescent model, making it difficult to assess fit to the coalescent model. Here we consider this problem from an analytic perspective by computing the probability of sampling two identical gene tree topologies, as well as the probability distribution of the pairwise Robinson-Foulds distance (hereafter “RF distance”, Robinson and Foulds, 1979) among possible gene trees, for model species trees with three to eight taxa.

Second, note that a well-known example of incongruent gene trees is the commonly observed conflict between mitochondrial and nuclear phylogenies (Ferris *et al.*, 1983; Moore, 1995; Sota and Vogler, 2001; Shaw, 2002; McCracken and Sorenson, 2005; Leaché and McGuire, 2006; Robertson *et al.*, 2006; Peters *et al.*, 2007; Good *et al.*, 2008). Conflicting mitochondrial and nuclear gene trees are often attributed to the low mutation rate of nuclear DNA sequences, hybridization or other horizontal gene transfer, and natural selection (Avise *et al.*, 1987; Rand, 2001; Sanderson and Shaffer, 2002; Funk and Omland, 2003; Ballard and Whitlock, 2004; Ballard and Rand, 2005; Spinks and Shaffer, 2009; Roos *et al.*, 2011), while the effect of ILS is often overlooked. Incongruence among gene trees is also extensively observed in other phylogenetic settings, such as conflicting plastid and nuclear gene trees, or different gene trees among nuclear genes (Cranston *et al.*, 2009; Moyer *et al.*, 2009; Bell and Hyvönen, 2010). Our computation of the expected extent of pairwise gene tree incongruence for specific examples serves to highlight the important role that ILS is likely playing in these empirical observations more generally.

Third, we consider gene tree incongruence as a potential source of bias when mapping character traits onto phylogenetic trees. In this setting, it is common to
use a single tree (normally an estimated species tree) as the phylogenetic framework onto which character traits (e.g., nucleotides, morphological traits, behavioral traits, etc.) are mapped. For example, Hahn and Nakhleh (2015) recently discussed the risk of ignoring variation in gene tree topologies and mapping characters onto a single representation of the species tree. Quantifying the extent to which the two underlying phylogenies may vary using the RF distance for specific examples, as we have done here, gives insights into the extent of the potential bias in mapping character traits onto a fixed phylogeny.

Fourth, we note that a similar issue may exist in phylogenetic comparative studies. A number of comparative studies look for correlation in two or more traits after adjusting for a single species tree. Often this single tree is a well-resolved species tree that is believed to be a reliable estimate of the species-level evolutionary relationships (Wiegmann et al., 2009; Misof et al., 2014). It is clear, however, that each of the traits under consideration may have its own gene tree, and these gene trees may vary substantially from the species tree. Although the uncertainty in phylogenetic relationships has been taken into account in several approaches (Richman and Price, 1992; Huelsenbeck et al., 2000; Lutzoni et al., 2001; Pagel and Lutzoni, 2002; Huelsenbeck and Rannala, 2003; Pagel et al., 2004), most of them focus on a lack of resolution in the overall species tree estimate (Hahn and Nakhleh, 2015) or gene tree estimation error (Simmons et al., 2016), rather than on genuine variation in the underlying histories. Again, quantification of the extent of incongruence can give insights into how robust comparative procedures might be to gene tree variation.

We next briefly describe our computational methods for assessing the extent of gene tree incongruence under the coalescent model, and then provide our results.
While we consider only small trees here (in the range of three to eight taxa), these examples serve to highlight the fact that gene tree incongruence is widespread, even when the number of taxa under consideration is not large. We discuss the implications of our findings in the Discussion.

2.2 Methods

In our study, the probability that two topologically identical gene trees are observed from the same species tree under the coalescent model is computed assuming no recombination within the two loci, and free recombination between the two loci. We further assume that the only evolutionary process generating discord between the two gene trees is the coalescent process (i.e., there is no hybridization or other horizontal gene transfer, and no gene duplication/loss). We consider the case in which one gene lineage is sampled for each species – when additional samples within a species are sampled, the situation will be even more complex in the sense that there is the possibility of even greater discord between gene trees. Varying population sizes and branch lengths of the species tree are examined to study their effect on congruence among gene trees. Furthermore, the probability distribution of pairwise RF distances between the two gene trees under the coalescent model is computed. Note that the expected gene tree congruence and the distribution of RF distances are computed based on the study of Degnan and Salter (2005). In their paper, they derived a method for computing the distribution of gene tree topologies given a bifurcating species tree with an arbitrary number of taxa, when one gene is sampled for each species. This method was implemented in the computer program COAL. When a given species tree (with branch lengths in coalescent units) is input into COAL, the probability of each
possible gene tree topology will be calculated. All calculations presented here are carried out using COAL (Degnan and Salter, 2005).

**Gene tree incongruence for species trees with varying numbers of taxa**

Completely asymmetric species trees and selected symmetric species trees with up to eight taxa are used to compute the probability of observing two topologically identical gene trees. We select three different levels for the internal branch lengths (time between one speciation event to the following speciation event) for each species tree. Species trees with all internal branch lengths equal to 2.0 (all branch lengths in coalescent units) are labeled “Long species trees”, species trees with all internal branch lengths equal to 1.0 are labeled “Medium species trees”, and species trees with all internal branch lengths equal to 0.5 are labeled “Short species trees”. Under the coalescent model, it is expected that gene trees sampled from species trees with “long” internal branch lengths will have little ILS, while species trees with “short” internal branch lengths are expected to generate gene trees with extensive ILS. It is impossible to explore the space of all species trees exhaustively, even for small numbers of taxa, because there are infinitely many choices for internal branch lengths. Our choice of “long”, “medium” and “short” settings here is meant to span the range of possibilities. The sum of squares of all possible gene tree probabilities gives the total probability of sampling two topologically identical gene trees for a given species tree.

### 2.2.1 Gene tree incongruence probability for species trees with varying internal branch lengths

To examine the effect of different branch lengths on the extent of gene tree incongruence, we use species trees of four taxa and five taxa. For the four-taxon case, the
species tree is denoted as \(((A : T_0, B : T_0) : T_1, C : (T_0 + T_1)) : T_2, D : (T_0 + T_1 + T_2))\),
where \(T_1\) and \(T_2\) are the two internal branch lengths. Note that the value of \(T_0\) does
not affect the probability of gene trees that are embedded in the species tree, because
only one individual is sampled in each species. In our calculations, we use 100 equally
spaced values from 0.0 to 5.0 for \(T_1\) and \(T_2\). For the five taxa case, the species tree
is denoted as \((((((A : T_0, B : T_0) : T_1, C : (T_0 + T_1)) : T_2, D : (T_0 + T_1 + T_2)) : T_3, E : (T_0 + T_1 + T_2 + T_3))\). Three levels (0.1, 1.0, and 5.0) are selected for \(T_3\), and 100
equally spaced values from 0.0 to 5.0 are used for \(T_1\) and \(T_2\), similar to the four-taxon
case. The probability of getting two topologically identical gene trees for each choice
of branch lengths is calculated as described in the last section.

2.2.2 Distribution of pairwise RF distance among gene trees

It is also of interest to measure the pairwise distance between a pair of randomly
sampled gene trees. We use the RF distance to measure this and consider the distribu-
tion of the pairwise RF distances among gene trees under the same species tree. The
pairwise RF distances between pairs of gene trees is measured for five-taxon species
trees and for seven-taxon species trees with varying internal branch lengths, to pro-
vide an example of how this distribution changes as both internal branch lengths and
number of taxa vary.

The five-taxon species trees have the format \((((((A : T_0, B : T_0) : T_1, C : (T_0 + T_1)) : T_2, D : (T_0 + T_1 + T_2)) : T_3, E : (T_0 + T_1 + T_2 + T_3))\). One of the three internal branch
lengths \(T_1\), \(T_2\), and \(T_3\) is varied from 0.5 to 5.0, while the other two branch lengths, as
well as \(T_0\), are fixed at 1.0. Both rooted and unrooted RF distances are calculated by
the RF.dist function given in the phangorn (Steel and Penny, 1993) package in R (R
The pairwise RF distance distribution is computed by summing up the probabilities (weighted by the gene tree probabilities) of getting a particular RF distance.

The seven-taxon species trees have an increased computational complexity, because there are 135,135 possible gene trees in total. Thus, we select the seven-taxa yeast species tree reported by Rokas et al. (2003) as the model species tree. To test the species tree with varying branch lengths, the estimated branch lengths from Fan and Kubatko (2011) are scaled by 0.8-, 1.0-, and 1.2-fold. The distribution of pairwise RF distances is computed in the same way as in the five-taxon case.

2.3 Results

2.3.1 Gene tree incongruence increases with an increasing number of taxa

The probability of independently drawing two gene trees that match from asymmetric species trees with varying numbers of taxa is shown in Figure 3(a), while Figure 3(b) represents the probability of getting matching gene trees from symmetric species trees. In both panels of Figure 3, three series of species trees with long internal branch lengths (2.0, the dotted line), medium internal branch lengths (1.0, the solid line) and short internal branch lengths (0.5, the dashed line) are used to calculate the gene tree incongruence. The results are consistent with our expectation that the probability of getting two topologically identical gene trees decreases as the number of taxa increases, and it is clear that this probability will become even smaller when there are more than eight taxa in the species tree. Notably, the probability that two gene trees are topologically identical is surprisingly low for only eight taxa with both medium and short internal branch lengths, illustrating that two gene trees will rarely
agree with each other for a large species tree (with more than eight taxa) when the internal branch lengths are not long enough. Even for the species trees with long internal branch lengths (2.0), the probability of sampling two topologically-matched gene trees is still below 40%. The results shown in Figure 3 also demonstrate that species trees with shorter branch lengths are more likely to generate incongruent gene trees, when the number of taxa is fixed. Even for a three-taxon species tree, it is still quite possible that two topologically different gene trees will be sampled (e.g., this happens with 56.4% of the time for the three-taxon species tree with “short” internal branch lengths (0.5)).

2.3.2 Gene tree incongruence decreases with increasing branch lengths of the species tree

To further explore the effect of branch lengths on the extent of gene tree incongruence, we use four-taxon and five-taxon asymmetric species trees with varying branch lengths. There are two internal branch lengths in the four-taxon species trees, and each is examined at 100 equally spaced values from 0.0 to 5.0 (Figure 4(a)). Unless both internal branches, $T_1$ and $T_2$, are long enough (larger than 2.0), it is very likely (prob. > 40%) that two topologically different gene trees will be sampled. It is highly possible (prob. > 80%) that the two gene trees match each other when both $T_1$ and $T_2$ exceed 3.0, which is a considerably long interval between speciation times. Notably, short speciation times will create more evenly distributed gene tree topologies, therefore increasing gene tree incongruence. Similar gene tree matching probability patterns appear in the five-taxon species tree cases (Figure 4(b)-(d)). We examine three different levels of $T_3$ (0.1, 1.0 and 5.0), and the overall probability that two sampled gene trees match increases with an increasing value of $T_3$. It is not
Figure 3: Gene tree congruence for species trees with different numbers of taxa. The x-axis shows the number of taxa in the completely asymmetric species tree, and the y-axis shows the probability of sampling two topologically identical gene trees. Species trees with long internal branch lengths (2.0) are denoted with the dotted line. Species trees with medium internal branch lengths (1.0) are denoted with the solid line. The dashed line denotes the species trees with short internal branch lengths (0.5). (a) Completely asymmetric species trees with species number varying from three to eight. (b) Selected symmetric species trees. Let x denotes the internal branch lengths (0.5, 1.0, or 2.0), the symmetric species trees are: Three taxa: ((A:x, B:x):x, C:2x); Four taxa: ((A:x, B:x):x, (C:x, D:x):x); Five taxa: (((A:x, B:x):x, C:2x):x, (D:2x, E:2x):x); Six taxa: ((((A:x, B:x):x, C:2x):x, (D:x, E:x):x, F:2x):x); Seven taxa: (((A:x, B:x):x, (C:x, D:x)):x, ((E:x, F:x):x, G:2x):x); Eight taxa: (((A:x, B:x):x, (C:x, D:x)):x, ((E:x, F:x):x, (G:x, H:x)):x).
surprising that with a short branch length of $T_3$ (less than 0.5), it is not possible for the probability that two gene trees match to exceed 60%. From Figure 4, it is clear that there are positive correlations between the extent of gene tree congruence and any internal branch lengths.

2.3.3 Distribution of RF distance between two gene trees

The rooted and unrooted pairwise RF distance distributions for gene trees sampled from several five-taxon species trees are shown in Figure 5. For a five-taxon species tree, the largest rooted RF distance is 6 and the largest unrooted RF distance is 4. When the RF distance is 0, the two gene trees have identical topologies. The effect of each internal branch length is examined by fixing the other branch lengths. It is quite reasonable that rooted gene trees (Figure 5 (a), (c), (e)) have a lower probability of matching each other or having similar topologies than the unrooted gene trees (Figure 5 (b), (d), (f)). However, even for the unrooted gene trees, the probability of sampling different gene trees exceeds 40% in all of the species trees we examined. The distributions of RF distance shown in Figure 5 also clearly indicate that two gene trees are more likely to agree with each other when the interval between speciation events is longer (Figure 5 (a)-(e)). The only exception is the unrooted RF distance of the species tree with fixed $T_1$ and $T_2$ (Figure 5 (f)). When $T_3$ varies from 0.5 to 5.0, the distribution of unrooted RF distance maintains the same. This phenomenon is not difficult to explain. As the most ancient internal branch, the length of $T_3$ only affects the root position of the gene trees, thus it does not change the topology of the unrooted gene trees.
Figure 4: Gene tree congruence for four-taxon and five-taxon species trees with varying internal branch lengths. The internal branch lengths $T_1$ and $T_2$ are denoted on the x-axis and the y-axis, respectively. The gene tree matching probability (varying from 0 to 1) is divided into 40 levels as shown in the color legend. (a) Four-taxon species tree 

$$(((A : T_0, B : T_0) : T_1, C : (T_0 + T_1)) : T_2, D : (T_0 + T_1 + T_2)),$$

where 100 equally spaced values from 0.0 to 5.0 for $T_1$ and $T_2$ are examined. (b) - (d) Five-taxon species tree 

$$(((A : T_0, B : T_0) : T_1, C : (T_0 + T_1)) : T_2, D : (T_0 + T_1 + T_2)) : T_3, E : (T_0 + T_1 + T_2 + T_3)).$$

Three levels of $T_3$ (0.1, 1.0, and 5.0) are examined. The other two branch lengths $T_1$ and $T_2$ are set as in the four-taxon case.
Figure 5: Distribution of pairwise RF distances among gene trees for the five-taxon species tree: (((((A : T₀, B : T₀) : T₁, C : (T₀ + T₁)) : T₂, D : (T₀ + T₁ + T₂)) : T₃, E : (T₀ + T₁ + T₂ + T₃))). One of the three internal branch lengths T₁ ((a), (b)), T₂ ((c), (d)), and T₃ ((e), (f)) varies from 0.5 to 5.0, while the other two branch lengths, as well as T₀, are fixed at 1.0. The varying internal branch length is denoted on the x-axis, and the probabilities of observing the RF distance for two sampled gene trees are denoted on the y-axis with different colors, as shown in the color legend. The left column ((a), (c), (e)) shows the rooted RF distances, while the right column ((b), (d), (f)) shows the unrooted RF distances.
Pairwise RF distances among gene trees are also measured in a seven-taxa yeast species tree (Rokas et al., 2003; Fan and Kubatko, 2011). We scaled the estimated branch lengths of this yeast species tree (Fan and Kubatko, 2011) by 0.8-, 1.0-, and 1.2-fold to assess the effect of branch length on the distribution of RF distances (Figure 6). Again, the branch lengths play an important role in the distribution of pairwise RF distances. Considering RF distances less than or equal to 4 as “similar”, and RF distances larger than 6 as “different”, the probability of getting “similar” gene trees is consistently decreasing with decreasing internal branch lengths. Conversely, it is more likely to get “different” gene trees when the branch lengths are shorter. Notably, in all three of the species trees, the probability of getting two “similar” gene trees that have a rooted RF distance less or equal than 4, or have an unrooted RF distance less or equal than 2 is smaller than or around 50%. Furthermore, the exact matching probability of two gene trees is considerably small (below 5% for the rooted gene trees and around 10% for the unrooted gene trees), which indicates that conflicting gene trees due to incomplete lineage sorting are common.

2.4 Discussion

As mentioned in the Introduction, coalescent theory is now extensively used in phylogenetic species tree inference, and the potential for discordance between gene trees and the species tree is well-studied (Degnan and Salter, 2005; Degnan and Rosenberg, 2006; Pollard et al., 2006; Liu and Pearl, 2007; Degnan and Rosenberg, 2009; Edwards et al., 2016). However, the importance of considering incongruent gene trees in studies that use phylogenies, but for which phylogeny estimation is not the ultimate goal, is rarely highlighted. In the settings of character mapping and
<table>
<thead>
<tr>
<th>RF</th>
<th>0.8</th>
<th>1</th>
<th>1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0180</td>
<td>0.0310</td>
<td>0.0473</td>
</tr>
<tr>
<td>2</td>
<td>0.0882</td>
<td>0.1354</td>
<td>0.1847</td>
</tr>
<tr>
<td>4</td>
<td>0.2232</td>
<td>0.2868</td>
<td>0.3316</td>
</tr>
<tr>
<td>6</td>
<td>0.3193</td>
<td>0.3213</td>
<td>0.2970</td>
</tr>
<tr>
<td>8</td>
<td>0.2659</td>
<td>0.1873</td>
<td>0.1229</td>
</tr>
<tr>
<td>10</td>
<td>0.0854</td>
<td>0.0383</td>
<td>0.0165</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RF</th>
<th>0.8</th>
<th>1</th>
<th>1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0505</td>
<td>0.0789</td>
<td>0.1111</td>
</tr>
<tr>
<td>2</td>
<td>0.1771</td>
<td>0.2408</td>
<td>0.2985</td>
</tr>
<tr>
<td>4</td>
<td>0.3547</td>
<td>0.3933</td>
<td>0.4004</td>
</tr>
<tr>
<td>6</td>
<td>0.3192</td>
<td>0.2403</td>
<td>0.1686</td>
</tr>
<tr>
<td>8</td>
<td>0.0985</td>
<td>0.0467</td>
<td>0.0214</td>
</tr>
</tbody>
</table>

Figure 6: Probability distribution of pairwise RF distances between gene trees in a seven-taxon species tree. The branch lengths are scaled by 0.8-, 1.0-, and 1.2-fold (headings of column 2 – 4) to calculate the probability of the RF distance. The headings “Rooted” and “Unrooted” refer to the RF distance between rooted gene trees and unrooted gene trees, respectively. Entries under the title “RF” denote the possible values of the RF distance.
phylogenetic comparative studies, for example, failure to consider gene tree incongruence generated by ILS has the potential to cause systematic biases in the results. To examine this issue in more detail, we quantified the extent of incongruence expected between pairs of gene trees sampled from a fixed species tree in two ways. First, we compute the probability that two sampled gene trees match topologically for species trees with three to eight taxa and varying branch lengths, and second, we compute the probability distribution of RF distances for specific species trees with four, five, and seven taxa. Our major finding is to demonstrate that the probability that two sampled gene trees match can be appreciably low, even when the trees under consideration are not large. Our results highlight that this issue should be of concern in the use of methods based on gene trees, as it cannot be expected, even in studies with limited taxon sampling, that the phylogenies underlying the traits of interest will be topologically congruent.

A high level of gene-tree conflict is widely observed in empirical studies (Spinks and Shaffer, 2009; Betancur-R et al., 2013; Salichos and Rokas, 2013; Pyron et al., 2014), though causes other than ILS are sometimes also hypothesized, such as rate variation among genes, gene flow/hybridization, natural selection or potential estimation error in gene tree inference (Avise et al., 1987; Rand, 2001; Sanderson and Shaffer, 2002; Funk and Omland, 2003; Ballard and Whitlock, 2004; Ballard and Rand, 2005; Spinks and Shaffer, 2009; Simmons et al., 2016). These causes can be difficult to distinguish from the ILS based on the estimated gene trees (Yang, 2002; Leigh et al., 2008; Betancur-R et al., 2013). Our study suggests one way in which this could be assessed, at least for small problems, as follows. First, pairwise RF distances among gene trees in empirical studies can be measured by the method described in
Simmons et al. (2016). Then, by contrasting the observed pairwise RF distances among gene trees with those expected for certain parameter choices, it may be possible to assess whether ILS is the dominant cause of the gene tree incongruence. To aid in this, we provide a set of scripts that can be used with COAL (Degnan and Salter, 2005) and R to compute gene tree probabilities and RF distance distributions (https://github.com/tian52/2016Match/tree/master/Scripts). We also provide files that will compute the probability of sampling two matching gene trees from asymmetric species trees with three to eight taxa for a set of user-specified branch lengths.

If the observed and expected pairwise RF distances among gene trees are similar, ILS is very likely to be the cause of conflict among observed gene trees. However, if the gene trees are more congruent than the expected, it is possible that the genes cannot be assumed to freely recombine, or that the genes are experiencing selection or convergent evolution (Hahn and Nakhleh, 2015). On the other hand, when the gene trees are less congruent than the expected, rate variation among genes could be important in the phylogenetic inference. Gene flow/hybridization and potential estimation error may affect the gene tree incongruence in either direction, depending on the specific case. Overall, we suggest that researchers should use the expected gene tree incongruence to assess whether ILS is a reasonable explanation for the extent of observed gene tree incongruence before explaining the gene tree incongruence by other causes.

Given the expectation of high levels of gene tree incongruence for even small taxon samples, the question becomes how to adequately deal with such incongruence in procedures for which estimated gene trees provide a necessary framework for inference, such as the phylogenetic comparative method. Clearly, correctly dealing with the
possibility of gene tree incongruence is an extremely difficult task, and more research in this area is necessary, as stated in Hahn and Nakhleh (2015). A potential solution in the comparative method framework is to incorporate phylogenetic uncertainty in the underlying trees for all traits under consideration by integrating over all possible gene trees under a particular evolutionary model. However, this approach would incur a substantial computational cost, and it is likely that more efficient approaches would need to be developed for problems that are large in either the number of traits or the number of taxa (or both).

Despite the complications that arise when dealing with the possibility of gene tree incongruence due to ILS, some of which are highlighted here, we argue that overall gene tree incongruence should be considered as additional phylogenetic information, rather than as “error” or “insufficient data”. When gene tree incongruence arises from ILS alone, it is actually an “intrinsic part of phylogeny’s nature”, as stated by Maddison (1997).
Chapter 3: Distribution of gene tree histories under the coalescent model with gene flow

We propose a coalescent model for three species that allows gene flow between both pairs of sister populations. The model is designed for multilocus genomic sequence alignments, with one sequence sampled from each of the three species, and is formulated using a Markov chain representation that allows use of matrix exponentiation to compute analytical expressions for the probability density of coalescent history. The coalescent history distribution as well as the gene tree topology distribution under this coalescent model with gene flow are then calculated via numerical integration. We analyze the model to compare the distributions of gene tree topologies and coalescent histories for species trees with differing effective population sizes and gene flow rates. Our results suggest conditions under which the species tree and associated parameters are not identifiable from the gene tree topology distribution when gene flow is present, but indicate that the coalescent history distribution may identify the species tree and associated parameters. Thus, the coalescent history distribution can be used to infer parameters such as the ancestral effective population sizes and the rates of gene flow in a maximum likelihood (ML) framework. We conduct computer
simulations to evaluate the performance of our method in estimating these parameters, and we apply our method to an Afrotropical mosquito data set (Fontaine et al., 2015).

### 3.1 Introduction

In multi-locus phylogenetic studies, many different evolutionary factors can cause incongruence between gene trees and species trees (Maddison, 1997). Incomplete lineage sorting (ILS; also called deep coalescence) has long been recognized to be one of the major causes of incongruence in gene trees across a genome (Pamilo and Nei, 1988; Takahata, 1989). Another important factor leading to incongruence between gene trees and the species tree is gene flow between populations following speciation (Maddison, 1997; Leaché et al., 2013). With some exceptions (noted below), these two processes have been studied in isolation. When carrying out phylogenetic analyses for species for which a sufficient amount of time has passed since speciation, ignoring gene flow following speciation might not bias the resulting estimates. However, with the advent of large-scale genomic data sets that allow for the study of evolutionary relationships among closely-related populations or species, the necessity of simultaneously examining these factors is becoming increasingly apparent (Eckert and Carstens, 2008; Leaché et al., 2013; Huang et al., 2014). In particular, both gene flow and incomplete lineage sorting are expected to occur between recently-diverged taxa (Yu et al., 2011) and so it is necessary to incorporate these processes simultaneously into models used to analyze data from closely related species or populations.

Degnan and Salter (2005) derived the probability distribution of gene trees under the coalescent model in the absence of gene flow and provided a method for computing
this distribution that was implemented in their software, COAL. Wu (2012) provided a method of computation that was more efficient than the method of Degnan and Salter, and used this method to develop software for species tree estimation called STELLS. Although both methods model the possibility of incomplete lineage sorting using the coalescent without gene flow, the difference between the two computational approaches is in the method of enumerating possible scenarios that are consistent with a given gene tree under the model. Degnan and Salter’s approach used the concept of gene tree coalescent histories, which are gene tree topologies together with an assignment of coalescent events on the gene tree topology to specific intervals of the species tree. In contrast, Wu used ancestral configurations to carry out the computations, where an ancestral configuration can loosely be defined as an assignment of possible states of all lineages at nodes of the species tree. In this way, Wu (2012) for details).

The ability to compute gene tree probability distributions provided several important insights into the problem of multi-locus species tree estimation. Important among these was the realization that the gene tree topology with the highest probability need not match the species tree; such gene trees are called anomalous gene trees (Degnan and Salter, 2005; Degnan and Rosenberg, 2006; Degnan et al., 2012). More broadly, these studies led to the realization that the incomplete lineage sorting process could result in substantial variation in the evolutionary trees for individual genes, suggesting the importance of accounting for this process in inferring species-level phylogenies. Another important insight was that the gene tree topology probability distribution identifies both the species tree topology and the speciation times (in coalescent units) (Allman et al., 2011a), which implies that if this distribution were known exactly then the species tree that produced it would also be known. This
has led to the development of a collection of methods for inferring species trees from estimated gene trees (Than and Nakhleh, 2009; Liu et al., 2010; Fan and Kubatko, 2011; Wu, 2012; Mirarab et al., 2014; Bayzid et al., 2015).

Some models that incorporate both gene flow and incomplete lineage sorting jointly have also been proposed. For example, a model with incomplete lineage sorting and gene flow leading to hybrid speciation was introduced to estimate the relative parental contributions to the hybrid taxon (Meng and Kubatko, 2009) and to detect hybridization within the framework of the coalescent model (Kubatko, 2009; Gerard et al., 2011). Yu et al. (2012, 2013) proposed a model that establishes a phylogenetic network to compute the probability of gene tree topologies (Yu et al., 2012, 2013), with “horizontal” branches in the network representing gene flow or hybridization events. A maximum likelihood method for inferring reticulate evolutionary histories with the existence of ILS was established by Yu et al. (2014). Maximum pseudo-likelihood approaches have been used as a fast and simple way to infer phylogenetic networks under ILS (Yu and Nakhleh, 2015; Solís-Lemus and Ané, 2015). Furthermore, Wen et al. (2016) developed methods for Bayesian inference of reticulate phylogenies under the multispecies network coalescent (Wen et al., 2016).

Isolation-with-migration (IM) models (Hey and Nielsen, 2004; Hey, 2010) have also been used to model both population splitting and gene flow. Zhu and Yang (2012) recently used this basic model to characterize the genealogical process with both coalescence and migration. In particular, Zhu and Yang (2012) calculated the distribution of coalescent histories under an IM model with two closely related species subject to gene flow and an outgroup species, and used this probability distribution to analyze sequence data for three taxa. They used the model to obtain estimates
of relevant parameters and to develop a hypothesis test for gene flow in a maximum likelihood framework. Andersen et al. (2014) used the two-population IM model with an arbitrary number of lineages in each population and derived gene tree probability distributions under this model. They also developed procedures for inferring model parameters from sequence data in this setting.

Here we propose a model for three species (plus an outgroup species for rooting, not shown in Figure 3.1) that allows gene flow between both pairs of sister populations. We formulate our model using the Markov chain representation of Hobolth et al. (2011), which allows use of matrix exponentiation to compute analytical expressions for the probability density of coalescent history. We then use numerical integration to calculate the coalescent history distribution as well as the gene tree topology distribution under this coalescent model with gene flow. Our work provides an extension to the work of Degnan and Salter (2005) that assumed no gene flow following speciation, in that it gives the entire probability distribution on gene tree topologies and coalescent histories under the coalescent model with gene flow between sister taxa. It thus enables consideration of identifiability of species tree parameters (extending the work of Allman et al. 2011 to the case in which gene flow following speciation is possible), insight into existing methods for inferring species trees that use only the gene tree topology distribution (e.g., (Liu et al., 2010; Degnan et al., 2009; Wu, 2012; Mirarab et al., 2014)), and evaluation of which features of empirical data might be important in inference settings in which gene flow is present (e.g., (Allman et al., 2011a, 2013) ).

One of our main findings is that, in contrast to the situation in the absence of gene flow (i.e., Allman et al. 2011), the species tree is not identifiable from the
gene tree topology distribution when gene flow is present. However, the coalescent
history distribution may identify the species tree topology. The finding that species
tree is not identifiable from the gene tree topology distribution has implications for
the use of coalescent-based species tree inference procedures that are based solely
on the observed frequencies of gene tree topologies estimated from multilocus data,
such as MP-EST (Liu et al., 2010), STELLS (Wu, 2012), and ASTRAL (Mirarab
et al., 2014). We emphasize that our work does not consider the effects of mutation
(i.e., all of our computations deal with the probability distributions of gene trees,
gene tree topologies, and coalescent histories arising from the species tree under the
coalescent model with gene flow) rather than considering sequence data directly. It
is clear that the effects of mutation arising on the individual gene trees will further
complicate species-level phylogenetic inference (see, e.g., (Lanier et al., 2014; Huang
et al., 2010; Huang and Knowles, 2009; McCormack et al., 2009; Liu and Edwards,
2009) for similar issues that arise in the analogous case in the absence of gene flow).
Note also that some work in calculating likelihoods of various mutational patterns
under similar models can be found in Lohse et al. (2011).

To provide a link between our calculations and empirical phylogenetic inference,
we develop a method that uses our work to infer model parameters, such as effective
population sizes and the rates of gene flow, from the coalescent history distribution
in a maximum likelihood (ML) framework. We conduct computer simulations to
evaluate the performance of our method in estimating the model parameters. We
apply our method to an Afrotropical mosquito data set (provided by Fontaine et al.,
2015) to demonstrate an application of our method to the analysis of empirical data.
We note that this application is somewhat specialized, in that some information about
the evolutionary history of the mosquitos (for example, the species tree topology, including times of speciation events) is already available, and because the method uses estimated gene trees as the input for inference (see, e.g., (Knowles et al., 2012; McCormack et al., 2011), for discussion of using gene trees vs. the original sequence data for inference in similar settings). Thus, we see our method as complementary to existing methods that seek to estimate evolutionary model parameters in a coalescent framework.

3.2 Materials and Methods

3.2.1 A model for three species with gene flow

In this paper, we develop a model for three species with gene flow between both pairs of sister taxa. In our proposed model, the species tree $((A, B), C)$ depicts relationships among the species A, B and C (Figure 3.1). The two ancestral species are denoted AB and ABC. The time since speciation occurred between A and B is denoted $\tau_1$, and the time since the speciation event between AB and C is denoted $\tau_2$, where $\tau_1$ and $\tau_2$ are measured by the expected number of mutations per site. The genetic data that we will analyze contain multiple loci. For every locus, we assume that one sequence was sampled from each of the three species. It is assumed that there is no recombination within a locus, and free recombination among loci. There are three possibilities for the gene tree topology relating the three sampled sequences at a locus: $((A, B), C)$, $((B, C), A)$, and $((A, C), B)$. Probability distributions relating to the coalescent history can be derived using Markov chains based on the structured coalescent process, as in Hobolth et al. (2011).
Figure 7: The model species tree \(((A, B), C)\) for three species with gene flow between sister species. The speciation times are denoted as $\tau_1$ and $\tau_2$, respectively. $\theta_A$, $\theta_B$, $\theta_C$, $\theta_{AB}$, and $\theta_{ABC}$ are the coalescent rates for each extant or ancestral species. The rates of gene flow between each sister species are assumed to be equal in both directions. The gene flow rate between species A and B is $m_1$, and the gene flow rate between species C and the ancestral species AB is $m_2$. 
Figure 8: (a) Five possible gene tree histories with gene tree topology \(((A, B), C)\) (denoted by $G_1$). For these gene tree histories, the first (most recent) coalescent event always occurs between the lineages from species A and B. If the first coalescent event occurs before $\tau_1$, and the second coalescent event occurs between $\tau_1$ and $\tau_2$, the gene tree has history $G_1H_1$. If the first coalescent event occurs before $\tau_1$, but the second coalescent event occurs after $\tau_2$, the gene tree has history $G_1H_2$. The other three gene tree histories are denoted by $G_1H_3$, $G_1H_4$, and $G_1H_5$. (b) Three possible gene tree histories with gene tree topology \(((B, C), A)\) (denoted by $G_2$) and three possible gene tree histories with gene tree topology \(((A, C), B)\) (denoted by $G_3$, species names labeled in parentheses). For these gene tree histories, the first coalescent event occurs between the lineages from species B and C (for topology $G_2$) or species A and C (for topology $G_3$). If the first coalescent event occurs before $\tau_1$, and the second coalescent event occurs between $\tau_1$ and $\tau_2$, the gene tree has history $G_2H_1$ or $G_3H_1$. The other two gene tree histories are denoted as $G_2H_2/G_3H_2$ and $G_2H_3/G_3H_3$. 

43
In our model, we consider gene flow between sister species A and B, and between sister species AB and C. More specifically, for species A and B, gene flow between contemporaneous populations can occur during the time period from the present to time $\tau_1$; for species AB and C, gene flow can occur between times $\tau_1$ and $\tau_2$ (assume that there is no gene flow between species B and C, or between species A and C, after time $\tau_1$). Additionally, to simplify the calculations, we assume that the gene flow rate between sister species is the same in both directions. The parameters involved in the model are: $\theta_A$, $\theta_B$, $\theta_C$, $\theta_{AB}$, $\theta_{ABC}$, $m_1$, $m_2$, $\tau_1$, and $\tau_2$ (see Figure 3.1). Here $\theta_A = 4N_A\mu$, $\theta_B = 4N_B\mu$, $\theta_C = 4N_C\mu$, $\theta_{AB} = 4N_{AB}\mu$, $\theta_{ABC} = 4N_{ABC}\mu$, where $N_x$ refers to the effective population size in species $x$, and $\mu$ is the mutation rate per site. The parameters $m_1$ and $m_2$ are defined to be the gene flow rates between the sister species (Hobolth et al., 2011).

We use the term gene genealogy to include information for both the gene tree topology and the associated coalescent times (Degnan and Rosenberg, 2009). For a given species tree with known speciation times, we can classify gene genealogies into coalescent histories based on where coalescent events occur in relation to speciation times. We note that there are infinitely many gene genealogies for any number of taxa because the coalescent times associated with the genealogy are continuous parameters. However, there are finitely many coalescent histories for any species tree since there are a finite number of speciation intervals into which the coalescent times can be placed (Figure 8).

High rates of gene flow will generate more variation in coalescent histories. Under our model in Figure 3.1, every species tree topology can have eleven possible coalescent histories. We denote a genealogy with gene tree topology ((A, B), C) by G1, a
genealogy with gene tree ((B, C), A) by G2, and a genealogy with gene tree ((A, C), B) by G3. Within each of these there is variation in the times at which coalescent events occur and we denote the possibilities by $H_x$ where $x$ is an integer. As shown in Figure 8(a), G1H1 to G1H5 show five different coalescent histories consistent with topology ((A, B), C). In Figure 8(b), G2H1 to G2H3 show three coalescent histories consistent with topology ((B, C), A). G3H1 to G3H3 are shown in Figure 8(c). The speciation times are labeled as $\tau_1$ and $\tau_2$ in the figure, while the coalescent time $t_1$ (for the first nearest coalescent event from present) and $t_2$ (for the second nearest coalescent event from present) are not labeled in the figure.

Under our model, we extend the Markov chain formulation of Hobolth et al. (2011) to calculate the probability distribution of the coalescent histories, and implement the method in a C program COALGF Calculator (COALGF). We then use the distributions of gene tree topologies and coalescent histories to compare and estimate species trees with differing effective population sizes and gene flow rates. To evaluate our method for the analysis of phylogenetic data with gene flow, we use simulation studies and an Afrotropical mosquito data set (Fontaine et al., 2015). Note that all parameters in our model (Figure 3.1) are scaled in COALGF so that they are proportional to a selected $c_0 = 2/\theta_0$ to simplify the calculation. In the following results, we use $\theta_x, c_x, m_x,$ and $\tau_x$ as the original parameters, which are not scaled by $2/\theta_0$. For the scaled parameters, we use $C_x = c_x/c_0, M_x = m_x/c_0,$ and $T_x = \tau_x * c_0$ as the scaled coalescent parameters, gene flow parameters, and speciation times, respectively. More details can be found in the Theory section.
3.2.2 Maximum likelihood parameter estimation

Using the results from the Theory section, the exact probabilities for the eleven coalescent histories can be calculated for any species tree with three species. Thus, given a data set consisting of observations of coalescent histories, these data can be viewed as a sample from a multinomial distribution with eleven categories and with probabilities as derived in the previous section. The likelihood of the data can thus be used to obtain maximum likelihood estimates of the coalescent parameters. We use simulation to assess the performance in estimating these parameters.

Three simulation studies were carried out using the software ms (Hudson, 2002) and seq-gen (Rambaut and Grassly, 1997). The ms software was used to simulate gene trees directly (for the first simulation study), while seq-gen was used to simulate 500bp and 1000bp DNA sequences on the simulated gene trees under the JC69 model (Jukes and Cantor, 1969) for the second and third simulation studies, respectively. The DNA sequence data sets were analyzed with PAUP* (Swofford, 2003) to estimate gene trees under the maximum likelihood criterion. In each simulation study, we selected a varying number of loci (ranging from 50 to 100,000) to assess our model. All data were simulated under the fixed species tree ((A, B), C), with $\theta_A = \theta_B = \theta_C = \theta_{AB} = 0.005$, $m_1 = m_2 = 200$, $\tau_1 = 0.004$, and $\tau_2 = 0.006$, which were chosen based on Zhu and Yang (2012). After scaling by $\theta_0 = 0.005$, we have: $C_1 = C_2 = C_3 = C_4 = 1$, $M_1 = M_2 = 0.5$, $T_1 = 1.6$, and $T_2 = 2.4$.

For each simulated data set with $K$ loci, the frequency of the $x^{th}$ coalescent history was recorded as $k_x, x = 1, 2, \ldots, 11$. In order to estimate the model parameters, we consider two methods for searching parameter space to find the maximum likelihood estimate (MLE), both based on a grid search. The first assumes that $M_1 = M_2 = M$
and \( C_1 = C_2 = C_3 = C_4 = C \), with \( C \) varying from 0 to 2 (we used 200 equal spaced values), and \( M \) varying from 0.001 to 10 (we used 200 different values on the log scale). The other method assumes that \( C_1 = C_2 = C_3 = C_4 = 1 \), and varies both \( M_1 \) and \( M_2 \) from 0.001 to 10 (we used 200 different values on the log scale). Note that although we could consider \( M_1, M_2, \) and \( C \) at the same time, in the simulation study, we only considered two parameters at a time in order to reduce the computational burden and to run more replications. For the empirical data, these parameters are estimated together.

For both methods, 40,000 species trees were tested. For each species tree, the exact probability distribution of the 11 coalescent histories was calculated and the likelihood for each simulated data set was calculated. The method of deriving the likelihood for the simulated data set follows Wang and Hey (2010) and Zhu and Yang (2012). The parameters with the highest likelihood are the maximum likelihood estimates. We simulated 1,000 replications for each simulation condition, and computed the average and the standard deviation of the MLEs of the model parameters over these replicates for each simulation condition. A naive implementation of the method in R takes approximately 2 hours to compute the MLE for each simulation condition (1,000 replications).

3.3 Theory

3.3.1 The probability distribution of coalescent histories

In our proposed model (Figure 3.1), we divide the species tree into three time periods. The first goes from the present time to \( \tau_1 \), and three species (A, B, and C) exist during this time period; the second time period goes from \( \tau_1 \) to \( \tau_2 \), with
two species AB and C; and the last goes from time $\tau_2$ to infinity, with only one species, ABC. In each time period, we can use the structured coalescent to describe the genealogical process. Hobolth et al. (2011) introduced the method to compute the probability density of times of coalescent events through matrix exponentials, and gave the instantaneous rate matrix for two populations with gene flow. We extend this method to our three-species model with gene flow between both pairs of sister populations.

**The instantaneous rate matrix for each time period**

During the first time period, from the present to $\tau_1$, gene flow can occur only between species A and B. A genealogy for a sample that includes one individual from each species has five possible states, which we denote by aac, abc, bbc, ac, bc. In our notation, aac means that two sequences are in species A, and one is in species C; abc means that one sequence is in each species; ac means that one sequence is in species A and another is in species C (here a pair of sequences have coalesced); and so on. Note that during this time period, species C always has one lineage since there is neither gene flow nor the possibility of a coalescent event, while the ancestral populations to species A and B can experience gene flow and/or a coalescent event among the two lineages. The rates of transitions between the five states can be expressed as a $5 \times 5$ instantaneous rate matrix $Q_1$: 

$$Q_1 = \begin{pmatrix}
    aac & abc & bbc & ac & bc \\
    -2m_1 - c_1 & 2m_1 & 0 & c_1 & 0 \\
    m_1 & -2m_1 & m_1 & 0 & 0 \\
    0 & 2m_1 & -2m_1 - c_2 & 0 & c_2 \\
    0 & 0 & 0 & -m_1 & m_1 \\
    0 & 0 & 0 & m_1 & -m_1
\end{pmatrix}$$
In this matrix, the coalescent parameters are defined as \( c_1 = 2/\theta_A; \ c_2 = 2/\theta_B; \ c_3 = 2/\theta_C; \ c_4 = 2/\theta_{AB}; \ c_5 = 2/\theta_{ABC}. \) We use \( e^Q \) to denote the matrix exponential \( e^Q = \sum_{i=0}^{\infty} \frac{Q^i}{i!}. \) The \((j,k)^{th}\) entry of \( e^Q \) is denoted as \((e^Q)_{jk}\).

Following Hobolth et al. (2011), we note that the probability density of a coalescent event at time \( t_1 \) during the time period from the present time to \( \tau_1 \) is

\[
f(t_1) = c_1(e^{Q_{11}t_1})_{21} + c_2(e^{Q_{11}t_1})_{23}, \quad t_1 < \tau_1.
\] (3.1)

When the first coalescent event occurs before time \( \tau_1 \) \((t_1 < \tau_1)\), a matrix \( Q_2 \) can be used to compute the probability density for the second coalescent event at time \( t_2 \), as follows. First denote species AB as population d, and species ABC as population e. The five possible genealogy states starting at time \( \tau_1 \) now become dd, dc, cc, d, c, and the associated rate matrix is

\[
Q_2 = \begin{pmatrix}
    & dd & dc & cc & d & c \\
    dd & -2m_2 - c_4 & 2m_2 & 0 & c_4 & 0 \\
dc & m_2 & -2m_2 & m_2 & 0 & 0 \\
cc & 0 & 2m_2 & -2m_2 - c_3 & 0 & c_3 \\
d & 0 & 0 & 0 & -m_2 & m_2 \\
c & 0 & 0 & 0 & m_2 & -m_2
\end{pmatrix}
\]

During the time period from \( \tau_1 \) to \( \tau_2 \), the probability density for the coalescent event at time \( t_2 < \tau_2 \) is

\[
f(t_2) = c_4(e^{Q_2(t_2-\tau_1)})_{21} + c_3(e^{Q_2(t_2-\tau_1)})_{23}, \quad t_2 < \tau_2.
\] (3.2)

The system becomes more complicated if the first coalescent event occurs after time \( \tau_1 \) \((t_1 > \tau_1)\). In that case, three lineages exist after time \( \tau_1 \) and all of them can migrate between species AB and C. A 13 \times 13 rate matrix \( Q_3 \) can be built to calculate the gene tree density. We label the state of each lineage sequentially. For
instance, ddc refers to the first two lineages being in population d (species AB), while
the third lineage is in population c (species C). There are 13 possible states, as shown
in matrix Q3,

\[
Q_3 = \begin{pmatrix}
\text{ddd} & \text{ddc} & \text{dcd} & \text{cdd} & \text{cdc} & \text{ccd} & \text{ccc} & \text{dd} & \text{dc} & \text{cc} & \text{d} & \text{c} \\
- & m_2 & m_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
m_2 & - & - & 0 & m_2 & 0 & m_2 & 0 & 0 & c_4 & 0 & 0 \\
m_2 & 0 & 0 & - & 0 & m_2 & 0 & m_2 & 0 & c_4 & 0 & 0 \\
m_2 & 0 & 0 & - & 0 & m_2 & 0 & m_2 & 0 & 0 & 0 & 0 \\
m_2 & 0 & 0 & - & 0 & m_2 & 0 & m_2 & 0 & c_3 & 0 & 0 \\
m_2 & 0 & 0 & - & 0 & m_2 & 0 & m_2 & 0 & 0 & 0 & 0 \\
m_2 & 0 & 0 & - & 0 & m_2 & 0 & m_2 & 0 & 0 & 0 & 0 \\
m_2 & 0 & 0 & - & 0 & m_2 & 0 & m_2 & 0 & 3c_3 & 0 & 0 \\
m_2 & 0 & 0 & - & 0 & m_2 & 0 & m_2 & 0 & 3c_4 & 0 & 0 \\
m_2 & 0 & 0 & - & 0 & m_2 & 0 & m_2 & 0 & 0 & c_4 & 0 \\
m_2 & 0 & 0 & - & 0 & m_2 & 0 & m_2 & 0 & 0 & 0 & 0 \\
m_2 & 0 & 0 & - & 0 & m_2 & 0 & m_2 & 0 & c_3 & 0 & 0 \\
m_2 & 0 & 0 & - & 0 & m_2 & 0 & m_2 & 0 & 0 & 0 & 0 \\
m_2 & 0 & 0 & - & 0 & m_2 & 0 & m_2 & 0 & 0 & 0 & 0 \\
m_2 & 0 & 0 & - & 0 & m_2 & 0 & m_2 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

where the diagonal entries (the ‘- ’ above) are set to the negative sum of the corre-
sponding row.

In our model, we use Q3 for the time period between time \( \tau_1 \) and \( t_1 \) when the first
coalescent event occurs after time \( \tau_1 \) (\( t_1 > \tau_1 \)). At the beginning of this time period,
there are 3 lineages with state ddc. At the end of this time period, since gene flow can
occur between populations d and c, any state with two or three lineages is possible.

The probability distribution of the gene tree histories.

Using the results above, the probability distribution of the coalescent histories can
be calculated. Recall that there are three different gene tree topologies, denoted as
G1, G2, and G3 for topologies \(((A, B), C), ((B, C), A), \text{ and } ((A, C), B)\), respectively
(Figure 8). Because coalescent events can occur in different intervals on the species
tree, there are multiple coalescent histories that are consistent with each gene tree
topology. For example, gene tree topology G1 can result from 5 different coalescent histories, labelled as G1H1, G1H2, and so on (Figure 8a). Similarly, G2 and G3 are both consistent with 3 different coalescent histories. The probability of each coalescent history will be calculated separately using the Markov chain formulation above. We give example calculations for a few coalescent histories below. The remaining calculations are given in the Appendix.

For G1H1, the first coalescent event occurs between the present and time τ1, while the second coalescent event occurs between time τ1 and τ2, thus τ1 < τ1 < τ2 < τ2. To derive the joint density of the coalescent times for this coalescent history, we first consider the time range between the present and time τ1. Since there is no gene flow between species C and any other species during this time, we only need to calculate the probability that the two lineages in species A and B coalesce before time τ1. Due to the possibility of gene flow between species A and B, the coalescent event can happen in either species A or B. The whole process can be described as three lineages that start in state abc, and then right before the first coalescent event, the state is either aac or bbc. Thus from (3.1), we can derive

\[ f_{G1H1}(t_1) = c_1(e^{Q1t_1})_{21} + c_2(e^{Q1t_1})_{23}, \quad t_1 < \tau_1 \]  

(3.3)

Similarly, the second coalescent event occurs between time τ1 and τ2, and the process is assumed to start from state dc. Right before the second coalescent event, the state is dd or cc. From the previous function (3.2), we have

\[ f_{G1H1}(t_2) = c_4(e^{Q2(t_2-\tau_1)})_{21} + c_3(e^{Q2(t_2-\tau_1)})_{23}, \quad \tau_1 < t_2 < \tau_2 \]  

(3.4)

The joint distribution for both \( t_1 \) and \( t_2 \) is then...
\[ f_{G_1H_1}(t_1, t_2) = f_{G_1H_1}(t_1) \times f_{G_1H_1}(t_2) \]
\[ = [c_1(e^{Q_1 t_1})_{21} + c_2(e^{Q_1 t_1})_{23}] [c_4(e^{Q_2(t_2 - \tau_1)})_{21} \]
\[ + c_3(e^{Q_2(t_2 - \tau_1)})_{23}, 0 < t_1 < \tau_1 < t_2 < \tau_2. \]

To find the marginal probability of gene tree history \( G_1H_1 \), we integrate out the
gene tree coalescent times,

\[ P(G_1H_1) = \int_0^{\tau_1} [c_1(e^{Q_1 t_1})_{21} + c_2(e^{Q_1 t_1})_{23}] \int_{\tau_1}^{\tau_2} [c_4(e^{Q_2(t_2 - \tau_1)})_{21} + c_3(e^{Q_2(t_2 - \tau_1)})_{23}] \, dt_2. \]

(3.5)

For \( G_1H_2 \), the first coalescent event occurs between the present and time \( \tau_1 \), while
the second coalescent event occurs after time \( \tau_2 \), and thus \( t_1 < \tau_1 < \tau_2 < t_2 \). The only
difference in the calculations for \( G_1H_2 \) is that no coalescent event occurs between
time \( \tau_1 \) and \( \tau_2 \). The probability that the second coalescent event occurs before time \( \tau_2 \) is

\[ P(t_2 < \tau_2 | t_1 < \tau_1 < t_2) = \int_{\tau_1}^{\tau_2} [c_4(e^{Q_2(t_2 - \tau_1)})_{21} + c_3(e^{Q_2(t_2 - \tau_1)})_{23}] \, dt_2. \]

(3.7)

Thus, the probability that the second coalescent event occurs after time \( \tau_2 \) is

\[ P(t_2 > \tau_2 | t_1 < \tau_1 < t_2) = 1 - \int_{\tau_1}^{\tau_2} [c_4(e^{Q_2(t_2 - \tau_1)})_{21} + c_3(e^{Q_2(t_2)})_{23}] \, dt_2, \]

(3.8)

and

\[ P(G1H2) = P(t_1 < \tau_1)P(t_2 > \tau_2|t_1 < \tau_1 < t_2) \]
\[ = \int_0^{\tau_1} [c_1(e^{Q_1 t_1})_{21} + c_2(e^{Q_1 t_1})_{23}] \, dt_1 \]
\[ \left(1 - \int_{\tau_1}^{\tau_2} [c_4(e^{Q_2(t_2 - \tau_1)})_{21} + c_3(e^{Q_2(t_2 - \tau_1)})_{23}] \, dt_2 \right). \]

(3.9)
For G1H3, the first coalescent event occurs after time $\tau_1$, while the second coalescent event occurs before time $\tau_2$, and thus $\tau_1 < t_1 < t_2 < \tau_2$. From the present to time $\tau_1$, no coalescent event occurs, and the probability is

$$P(t_1 > \tau_1) = 1 - \int_0^{\tau_1} [c_1(e^{Q_{11}t_1})_{21} + c_2(e^{Q_{11}t_1})_{23}] \, dt_1. \quad (3.10)$$

After time $\tau_1$, things are more complicated. There are four distinct ways in which the two coalescent events can happen. In all four cases, the process starts in state ddc, which means that two lineages are in population d and one is in population c. The first case, denoted by G1H3C1, goes from state ddc to ddd, and then a coalescent event occurs and the state is dd. The final coalescent event can occur either from this state, or by changing to state cc. To model this, we use $Q_3$ to calculate the change from state ddc to ddd, and use $Q_2$ for the change from state dd to dd or cc. The joint density function is

$$f_{G1H3C1}(t_1, t_2) = c_4(e^{Q_{31}t_1})_{21}[c_4(e^{Q_{21}(t_2-\tau_1)})_{11} + c_3(e^{Q_{21}(t_2-\tau_1)})_{13}]. \quad (3.11)$$

Notice that in the above density function, there is no multiplier for $c_4$ in the first coalescent process. Although there are three possible lineage combinations that can coalesce at time $t_1$, only one of them (the lineages that come from species A and species B) will maintain the gene tree topology ((A, B), C).

Similarly, we can write the density functions for the other three possibilities. The second possibility has the sequence of states ddc - ddc - dc - dd/cc. The sequence shows the process that three lineages start at ddc, then just before the first coalescence they are still at status ddc, after the first coalescence they are at status dc, then just before the second coalescence they are either at dd or cc, and then the last coalescence
happens. Thus, the sequence of states ddc - ddc - dc - dd/cc has corresponding density function

\[ f_{G1H3C2}(t_1, t_2) = c_4(e^{Q_1 t_1})_{22}[c_4(e^{Q_2(t_2-t_1)})_{21} + c_3(e^{Q_2(t_2-t_1)})_{23}]. \quad (3.12) \]

The third possibility has the sequence of states ddc - ccc - cc - dd/cc, with corresponding density function

\[ f_{G1H3C3}(t_1, t_2) = c_3(e^{Q_1 t_1})_{28}[c_4(e^{Q_2(t_2-t_1)})_{31} + c_3(e^{Q_2(t_2-t_1)})_{33}]. \quad (3.13) \]

Finally, the fourth possibility has the sequence of states ddc - ccd - dc - dd/cc, with corresponding density function

\[ f_{G1H3C4}(t_1, t_2) = c_3(e^{Q_1 t_1})_{27}[c_4(e^{Q_2(t_2-t_1)})_{21} + c_3(e^{Q_2(t_2-t_1)})_{23}]. \quad (3.14) \]

Thus, the overall density function for \( t_1 \) and \( t_2 \) for history G1H3 is

\[ f_{G1H3}(t_1, t_2) = f_{G1H3C1}(t_1, t_2) + f_{G1H3C2}(t_1, t_2) + f_{G1H3C3}(t_1, t_2) + f_{G1H3C4}(t_1, t_2), \quad (3.15) \]

and the marginal probability of G1H3 is

\[ P(G1H3) = P(t_1 > \tau_1)[\int_{\tau_1}^{t_2} \int_{\tau_1}^{t_2-t_1} f(G1H3, t_1, t_2) \, dt_2 \, dt_1]. \quad (3.16) \]

The probabilities of all other gene tree histories can be calculated similarly. We give the details in the Appendix 1.

### 3.3.2 Implementation details and parameter scaling.

Calculating the probability for each coalescent history requires computing integrals or double integrals (for coalescent histories G1H3, G2H1 and G3H1). To do
this, we used Gaussian Quadrature for one-dimensional integration, and iterated it for two-dimensional integration. Gaussian Quadrature is a efficient way in numerical analysis as an approximation of the definite integral. The method is implemented in a C program COALGF that directly calculates the probabilities for all eleven coalescent histories as well as the three gene tree topologies. The required input parameters in COALGF are the coalescent rates $c_1, c_2, c_3,$ and $c_4$ (for population A, B, C, and AB, respectively); the gene flow rates $m_1$ and $m_2$ ($m_1$ for the gene flow rate between population A and B, $m_2$ for the gene flow rate between population AB and C; we assume equal rates of gene flow to and from sister species); and the speciation times $\tau_1$ and $\tau_2$.

To simplify the calculation, all parameters are scaled in COALGF so that they are proportional to a selected $c_0 = 2/\theta_0$. Consider the probability of the first gene tree history in (3.6). Using matrix $Q1$ as an example, note that for $c_0 = 2/\theta_0$ we can write

$$Q1 = c_0* \begin{pmatrix} aac & abc & bbc & ac & bc \\ aac & -2m_1/c_0 - c_1/c_0 & 2m_1/c_0 & 0 & c_1/c_0 \\ abc & m_1/c_0 & -2m_1/c_0 & m_1/c_0 & 0 \\ bbc & 0 & 2m_1/c_0 & -2m_1/c_0 - c_2/c_0 & 0 \\ ac & 0 & 0 & 0 & c_2/c_0 \\ bc & 0 & 0 & 0 & 0 \end{pmatrix}$$

Note that $Q1 = c_0 * Q1'$, where $Q1'$ is the new matrix above, with all coalescent rates and gene flow rates scaled by $c_0$, and with $t'_1 = t_1 * c_0$. Similarly we can scale all coalescent rates by dividing $c_0$, and all speciation times by multiplying $c_0$. 

55
3.4 Results

3.4.1 Gene flow between sister species produces different distributions of gene tree histories

We calculated the probability distribution of coalescent histories using the method described above under a set of species trees with different parameter values (Figure 9). It is clear that the sum of the probability of G1 and twice of the probability of G2/3 is equal to 1 under each species tree setting. The effect of gene flow on the distribution of coalescent histories is explored under two different conditions: all current and ancestral populations have equal effective population sizes (Figure 9(a)), and current and ancestral populations have unequal effective population sizes (Figure 9(b)).

When there is no gene flow between either pair of sister species (Figure 9(a) \(M_1 = 0, M_2 = 0\)), only three coalescent histories are possible (G1H4, G1H5, and G2H3/G3H3; see Figure 8), since the first coalescent event cannot occur before speciation time \(T_1\), and the second coalescent event cannot occur before speciation time \(T_2\). When gene flow can occur between species AB and C, the second coalescent event could also occur between speciation time \(T_1\) and \(T_2\), and thus G1H3, G2H1/G3H1 and G2H2/G3H2 (see Figure 8) will also have positive probability. In that case, topology G1 contains three different coalescent histories, while topologies G2/G3 also have three different coalescent histories (Figure 9(a) \(M_1 = 0, M_2 = 0.5\)). However, if there is no gene flow between species AB and C, but gene flow is possible between species A and B (Figure 9(a) \(M_1 = 0.5, M_2 = 0\)), again only one coalescent history (G2H3/G3H3) is possible for topologies G2/3, since the second coalescent event cannot occur before speciation time \(T_2\), but the first coalescent event can occur before
Figure 9: The probability distribution of the gene tree histories under species trees with scaled speciation times $T_1 = 1.6$, $T_2 = 2.4$ ($\tau_1 = 0.004$, $\tau_2 = 0.006$). Each gene tree history is denoted by a different color as shown in the figure. For example, in the first bar we see that 55% of the probability can be attributed to history 1 and 15% can be attributed to history 2 within topology 1. Histories are grouped according to their topology. The probability of topology G1 is shown by the height of the column labeled G1; the height of the column labeled G2/3 shows the equal probability of the topologies G2 and G3. Thus, $P(G1) + 2P(G2/3) = 1$. The two sets of scaled coalescent rates are $C_1 = C_2 = C_3 = C_4 = 1$ (scaled by $\theta_0 = 0.005$) in panel (a), and $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2$ (scaled by $\theta_0 = 0.005$) in panel (b). Each panel contains four cases of different gene flow rates: 1, no gene flow ($M_1 = M_2 = 0$); 2, no gene flow between species A and B ($M_1 = 0, M_2 = 0.5$); 3, no gene flow between species C and the ancient species AB ($M_1 = 0.5, M_2 = 0$); 4, equal rates of gene flow in both sister species ($M_1 = M_2 = 0.5$).
speciation time $T_1$. The gene tree topology G1 will still contain three coalescent histories, but these are different from the three that appear with $M_1 = 0, M_2 = 0.5$. When gene flow is possible between both pairs of sister species (species A & B and species AB & C), all of the coalescent histories in Figure 8 have positive probability (Figure 9(a) $M_1 = 0.5, M_2 = 0.5$).

While the coalescent history distribution shows a clear pattern when all effective population sizes are assumed to be equal (Figure 9(a)), the distribution changes when these parameters differ across the species tree. In Figure 9(b), we used a set of species trees with scaled coalescent parameters $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2$ along with all of the other parameter combinations in Figure 9(a). The ratio of the effective population sizes considered here was selected based on Burgess and Yang’s (2008) paper in which hominoid ancestral population sizes were estimated. With regard to which settings lead to positive probabilities associated with various coalescent histories, the patterns in Figure 9(a) and (b) are generally consistent. Except for the case in which $M_2 \neq 0$, coalescent histories G1H3 and G2H1/G3H1 have extremely small probabilities (Figure 9(a)(b) $M_1 = 0, M_2 = 0.5; M_1 = 0.5, M_2 = 0.5$). This result is not surprising, because the length of time between the two speciation times $T_1$ and $T_2$ is relatively small ($T_2 - T_1 = 0.8$). Coalescent histories G1H3 and G2H1/G3H1 only arise when both coalescent events occur between $T_1$ and $T_2$ (Figure 8). With a small time span between $T_1$ and $T_2$ as well as unequal effective population sizes, these probabilities become very small. If the time span between $T_1$ and $T_2$ is longer, as in Figure 16(b), $T_1 = 2$ and $T_2 = 4$, it is clear that all five coalescent histories show up in topology G1, and all three coalescent histories appear in G2/3, when $M_2 \neq 0$. Finally, through we observe a similar pattern for the coalescent history probabilities for the
species trees with equal or unequal population sizes, the magnitude of the probability of each coalescent history varies significantly. For example, in Figure 9(a), G1H4 is one of the dominant coalescent histories, while in Figure 9(b), the probability of G1H4 is much smaller and the probabilities of G1H5 and G2H3/G3H3 are dominant.

We also considered another set of speciation times, $T_1 = 2$ and $T_2 = 4$, which has a much longer time span between the two speciation events (Figure 16). Comparing Figure 9(a) and Figure 16(a) shows that with longer speciation times, the probabilities of coalescent histories G1H5 and G2H3/G3H3 clearly decrease because the increase in speciation times decreases the probability that both coalescent events occur before the earliest speciation time. Interestingly, if there is no gene flow between either pair of sister species, the longer speciation times will greatly decrease the probability of observing the “incorrect” gene tree topology (G2 or G3). The stated pattern of longer speciation times leading to a lower probability of the wrong topology is due to the decreased probability of deep coalescence that comes with longer speciation times. Gene flow between only species A and B, but not species AB and C, will lead to a similar distribution because the topology G2/G3 is still composed of just one coalescent history (G2H3/G3H3) and the probability of this coalescent history decreases with longer speciation times. However, if ancestral gene flow exists between species AB and C, the distribution of gene tree topologies does not change much, but the distribution of coalescent histories shows some clear changes (compare Figure 9(a)(b) and Figure 16(a)(b)).

We also considered a second level for the rate of gene flow. In Figure 16(c)-(f), the rate of gene flow was set to 2 when it was present (in contrast to the rate of 0.5 used in Figure 9 and Figure 16(a),(b)). For the same effective population sizes and
the same speciation times, we find that changing the rate of gene flow from 0.5 to 2 does not have a large effect on the distribution of coalescent histories. More extreme values of the rate of gene flow will be discussed in the following sections.

3.4.2 Variation in the gene tree history distribution as a function of the rate of gene flow

We considered the change in the probabilities of individual coalescent histories as a function of the gene flow rate when all other parameters were held constant. In each subplot of Figure 11 and Figures 17 - 20, $M_1$ is held constant (at four different levels $M_1 = 0.001, 0.5, 2, 20$, corresponding to the rows), while the value of $M_2$ changes from 0 to 1,000. The effective population sizes and speciation times both have two different levels. For the effective population sizes, one setting is that all effective population sizes are equal, thus the coalescent rates were set to $C_1 = C_2 = C_3 = C_4 = 1$ (Figure 11(a)-(i), Figures 17, 18), and the other setting is a species tree with unequal effective population sizes, $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2$ (Figure 11(j)-(l), Figures 19, 20). For the speciation times, a set of relatively shorter speciation times [$T_1 = 1.6; T_2 = 2.4$ (Figure 11(a)-(f),(j)-(l), Figures 17, 19)] and a set of longer speciation times [$T_1 = 2; T_2 = 4$ (Figure 11(g)-(i), Figures 18, 20)] are used. Figure 10 and Figures 21 - 24 have similar parameter settings, but in these figures, $M_2$ is held constant at four different levels, while the value of $M_1$ is changed from 0 to 1,000. In all of these figures that plot the distribution of coalescent histories against the rate of gene flow, the first column shows the distribution of the five coalescent histories with topology $G_1=((A,B),C)$, the second column shows the distribution of the three coalescent histories with the other two topologies $G_2/G_3$, and the last column shows the distribution of the three possible gene tree topologies ($G_1$, $G_2$, $G_3$).
and G3). It is not surprising that the probabilities of the topologies G2 and G3 are always equal, as seen in the last column, because when the gene flow rate from species A to species B is identical to the reverse direction, we have \( P(G2H1) = P(G3H1) \), \( P(G2H2) = P(G3H2) \), and \( P(G2H3) = P(G3H3) \).

When there is no gene flow between species A and B (i.e., \( M_1 = 0.001 \); Figure 11(a)-(c)) or between species AB and C (i.e., \( M_2 = 0.001 \); Figure 10(a)-(c)), different coalescent histories will have positive probability compared with the case in which gene flow is present. For example, comparing Figure 11(a)-(c) and (d)-(f), we see that some coalescent histories have positive probability only when there is gene flow (G1H1 and G1H2). Comparing Figure 10(a)-(c) and (d)-(f), we find the same pattern, but with different coalescent histories (G1H1 and G1H3). Notably, determining whether or not gene flow occurred based on the frequencies of gene tree topologies would be difficult, especially when the rate of gene flow is not large (Figure 11(c),(f); Figure 10(c),(f)). Another interesting result is that if the rate of gene flow is held constant for one pair of sister species, the coalescent history distribution can change completely as the other gene flow rate changes from a small value to a large value (Figure 11 and Figure 10). Is it not surprising since the increasing gene flow rate will actually break the species boundary, thus changes the coalescent history distribution completely. These results suggest that the distribution of coalescent histories depends highly on the magnitude of gene flow.

In addition to gene flow, two other factors may affect the distribution of coalescent histories. The first factor is the speciation time (Figure 11(d)-(f),(g)-(i)). When \( M_2 \) is low (less than 0.1), the distributions of coalescent histories under different speciation times show a very similar pattern with slightly different values. However, when \( M_2 \) is
Figure 10: Probability distributions of the coalescent histories for the model of three species with gene flow between sister species. The probabilities of each coalescent history (y-axis) were plotted against the gene flow rate between species A and species B, $M_1$ (x-axis; shown on a log scale). The four sets of parameter values are $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 1.6$, $T_2 = 2.4$, $\theta_0 = 0.005$, $M_2 = 0.001$ for panels (a) - (c); $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 1.6$, $T_2 = 2.4$, $\theta_0 = 0.005$, $M_2 = 2$ for panels (d) - (f); $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.01$, $M_2 = 2$ for panels (g) - (i); and $C_1 = 1$, $C_2 = C_3 = 0.5$, $C_4 = 0.2$, $T_1 = 1.6$, $T_2 = 2.4$, $\theta_0 = 0.005$, $M_2 = 2$ for panels (j) - (l). PZ indicates population size.
Figure 10

(a) M2 = 0.001
Equal PZ
T1 = 1.6
T2 = 2.4

(b) M2 = 2
Equal PZ
T1 = 1.6
T2 = 2.4

(c) M2 = 2
Equal PZ
T1 = 2
T2 = 4

(d) M2 = 2
Unequal PZ
T1 = 1.6
T2 = 2.4

63
Figure 11: Probability distributions of the gene tree histories for the model of three species with gene flow between sister species. The probabilities of each gene tree history (y-axis) were plotted against the gene flow rate between species C and the ancient species AB, $M_2$ (x-axis; shown on a log scale). Panels (a), (d), (g), and (j) show the probabilities of the five gene tree histories (G1H1 - G1H5) with topology G1. Panels (b), (e), (h), and (k) show the probabilities of the three gene tree histories (G2H1/G3H1 - G2H3/G3H3) with topology G2 or G3. Panels (c), (f), (i), and (l) show the probabilities of the three gene tree topologies (G1, G2 and G3) with $P(G2) = P(G3)$, and $P(G1) + P(G2) + P(G3) = 1$. The four sets of parameter values are $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 1.6$, $T_2 = 2.4$, $\theta_0 = 0.005$, $M_1 = 0.001$ for panels (a) - (c); $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 1.6$, $T_2 = 2.4$, $\theta_0 = 0.005$, $M_1 = 2$ for panels (d) - (f); $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.01$, $M_1 = 2$ for panels (g) - (i); and $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2, T_1 = 1.6, T_2 = 2.4, \theta_0 = 0.005$, $M_1 = 2$ for panels (j) - (l).
Figure 11

(a) M₁ = 0.001
Equal PZ
T₁ = 1.6
T₂ = 2.4

(b) M₁ = 2
Equal PZ
T₁ = 1.6
T₂ = 2.4

(c) M₁ = 2
Unequal PZ
T₁ = 1.6
T₂ = 2.4

(d) M₁ = 0.001
Equal PZ
T₁ = 2
T₂ = 4

(e) M₁ = 2
Equal PZ
T₁ = 2
T₂ = 4

(f) M₁ = 2
Unequal PZ
T₁ = 2
T₂ = 4

(g) M₁ = 0.001
Equal PZ
T₁ = 1.6
T₂ = 4

(h) M₁ = 2
Equal PZ
T₁ = 1.6
T₂ = 4

(i) M₁ = 2
Unequal PZ
T₁ = 1.6
T₂ = 4

(j) M₁ = 0.001
Equal PZ
T₁ = 2
T₂ = 4

(k) M₁ = 2
Equal PZ
T₁ = 2
T₂ = 4

(l) M₁ = 2
Unequal PZ
T₁ = 2
T₂ = 4
larger, the distributions of coalescent histories with the two sets of different speciation
times have very different patterns. It is quite interesting to notice that the topology
distribution differs more when \( M_2 \) is small (less than 0.1), but becomes more similar
as \( M_2 \) becomes larger (Figure 11(f) and (i)). This effect is even larger in Figure
10(d)-(f) and (g)-(i), when \( M_2 \) is held constant and \( M_1 \) varies.

The other factor that may affect the distribution of coalescent histories is the
effective population size, which determines the rate at which coalescent events occur.
We observe that, for the two different sets of coalescent rates we considered (equal
effective population sizes \([C_1 = C_2 = C_3 = C_4 = 1]\) and unequal effective population
sizes \([C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2]\)), the gene tree topology distribution is affected
less by the change of effective population sizes than the coalescent history distribution
(Figure 11(d)-(f) and (j)-(l); Figure 10(d)-(f) and (j)-(l); Figure 17 - 24).

3.4.3 Different distributions of gene tree histories may share
an identical gene tree topology distribution

Under the typical coalescent model without the possibility of gene flow following
speciation, the distribution of gene tree topologies can be used to estimate the un-
rooted species tree topology and some information of branch lengths (Allman et al.,
2011a,b). Note that the branch lengths estimated here consider the combined effect
of population size and mutation rate. However, in the presence of gene flow, the gene
tree topology probabilities change, and we might ask whether this distribution alone
is sufficient to identify whether or not gene flow has occurred. Our overall finding is
that many different distributions of coalescent histories arising from different species
trees may share an identical gene tree topology distribution, which indicates that the
information about gene tree topology probabilities alone is not enough to estimate species trees in the presence of gene flow.

For example, Figure 12 shows eight species trees ($S_1 - S_8$) that all have the same set of scaled speciation times, $T_1 = 1.6$ and $T_2 = 2.4$. Trees $S_1 - S_4$ have equal effective population sizes with scaled coalescent parameters $C_1 = C_2 = C_3 = C_4 = 1$, while species trees $S_5 - S_8$ have unequal effective population sizes with scaled coalescent parameters $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2$. The eight species trees differ in the rates of gene flow, $M_1$ and $M_2$. For instance, $S_1$ shows the case of very low gene flow ($M_1 = M_2 = 0.001$), while $S_5$ shows the case that gene flow occurs mostly between species A and B ($M_1 = 0.798; M_2 = 0.001$). Different levels of gene flow are also reflected in this figure: $S_2$ has small gene flow rates ($M_1 = 0.5; M_2 = 0.544$); $S_8$ has medium gene flow rates ($M_1 = 2; M_2 = 5$); and $S_4$ has fairly large gene flow rates ($M_1 = 20; M_2 = 28$). As labeled in the figure, different colors show the probabilities of different coalescent histories. For each species tree, the probabilities of the eleven coalescent histories sum up to 1. The two solid black vertical lines in Figure 12 divide the total probability into the three probabilities corresponding to the gene tree topologies (from left to right, $G_1$, $G_2$, and $G_3$), which are identical for all eight cases. Under this identical topology distribution ($P(G_1) = 0.7, P(G_2) = P(G_3) = 0.15$), the distributions of coalescent histories are very different for these eight species trees.

Notably, these eight species trees are not the only species trees that share this particular gene tree topology distribution, and this gene tree topology distribution is not the only one which can be generated by multiple species trees. However, despite the implied non-identifiability of gene flow based only on the topology distribution, we note that the coalescent history distributions appear to be distinct in these cases.
Figure 12: Coalescent history distributions for eight species trees (rows labeled $S_1 - S_8$) with different coalescent rates and gene flow rates. The $x$-axis is the probability associated with individual coalescent histories (denoted by colors), and the total probability for each case sums up to 1. The two solid black lines indicate the probability of each of the three gene tree topologies, from left to right $P(G_1) = 0.7$, and $P(G_2) = P(G_3) = 0.15$. All eight species trees have scaled speciation times $T_1 = 1.6$, $T_2 = 2.4$ (scaled by $\theta_0 = 0.005$). The two sets of coalescent rates are $C_1 = C_2 = C_3 = C_4 = 1$ for species trees $S_1 - S_4$, and $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2$ for species trees $S_5 - S_8$. All species trees have different rates of gene flow: $S1: M_1 = M_2 = 0$; $S2: M_1 = 0.5, M_2 = 0.544$; $S3: M_1 = 2, M_2 = 2.23$; $S4: M_1 = 20, M_2 = 28$; $S5: M_1 = 0.796, M_2 = 0$; $S6: M_1 = 1.337, M_2 = 0.5$; $S7: M_1 = 1.894, M_2 = 2$; $S8: M_1 = 2, M_2 = 5$. 68
though we have not shown this formally. Because of this, the information provided by the distribution of coalescent histories can be used to estimate the parameters (effective population sizes and gene flow rates) in the coalescent model with gene flow.

3.4.4 Simulation studies for maximum likelihood parameter estimation

Based on these results, we developed a method to estimate the rates of coalescence and of gene flow along a fixed species phylogeny using maximum likelihood (see Materials and Methods for details). To assess the performance of our method, we carried out three simulation studies. In the first simulation study, we simulated gene trees directly with a varying number of loci ranging from 50 to 100,000. Gene trees were simulated under the fixed species tree ((A, B), C), with $\theta_A = \theta_B = \theta_C = \theta_{AB} = 0.005$, $m_1 = m_2 = 200$, $\tau_1 = 0.004$, and $\tau_2 = 0.006$, corresponding to scaled parameters $C_1 = C_2 = C_3 = C_4 = 1$, $M_1 = M_2 = 0.5$, $T_1 = 1.6$, and $T_2 = 2.4$. Assuming that $M_1 = M_2 = M$ and $C_1 = C_2 = C_3 = C_4 = C$, we evaluated the likelihood of 40,000 species trees (each differing in the values of $M$ and $C$ but with the topology and branch lengths fixed) to find the maximum likelihood estimates (MLEs) of $M$ and $C$. Figure 13 shows the results for one simulated data set in this simulation study for which 1000 gene trees were simulated with $C = 1$ and $M = 0.5$. It is clear that the MLEs $\hat{C} = 0.99$ and $\hat{M} = 0.468$, marked with a white ‘X’ in Figure 13, are close to the true values.

We repeated this simulation process 1,000 times for varying numbers of loci, and obtained the mean and standard deviation for the MLEs of $C$ and $M$ (Figure 14(a), top two plots). We see that the estimates of $C$ and $M$ appear to be generally
Figure 13: Contour plot for one simulated data set showing the log likelihood as a function of the scaled coalescent rate $C$ and the gene flow rate $M$, assuming that $C_1 = C_2 = C_3 = C_4 = C$ and $M_1 = M_2 = M$. DNA sequence data were simulated for 1000 loci for species tree $((A, B), C)$ with scaled speciation times $T_1 = 1.6$, $T_2 = 2.4$ (scaled by $\theta_0 = 0.005$), coalescent rates $C_1 = C_2 = C_3 = C_4 = 1$ and gene flow rates $M_1 = M_2 = 0.5$. The true scaled speciation times $T_1 = 1.6$, $T_2 = 2.4$ were used to identify different coalescent histories. The MLEs are $\hat{C} = 0.99$, and $\hat{M} = 0.468$, indicated by the white cross in the plot, with log likelihood $l = -2001.745$. The white dots in the figure are missing values caused by some numerical issues when preforming matrix exponential.
to simulate the data in all cases were scaled coalescent rates are fixed at their true values 1,000bp sequences simulated by ms and then seq-gen. The plots labeled "Estimated model with gene flow between both pairs of sister taxa. (a) Gene trees are directly Figure 14: Maximum likelihood estimates of the scaled coalescent rates and the gene flow rates obtained from the simulated data sets under the three species coalescent model with gene flow between both pairs of sister taxa. (a) Gene trees are directly simulated from the given species tree by ms; (b) Gene trees were estimated from 500bp sequences simulated by ms and then seq-gen; (c) Gene trees were estimated from 1,000bp sequences simulated by ms and then seq-gen. The plots labeled "Estimated C" and "Estimated M" on the y-axis refer to the MLEs of C and M under the assumption that $M_1 = M_2 = M$, and $C_1 = C_2 = C_3 = C_4 = C$. The plots labeled "Estimated $M_1$" and "Estimated $M_2$" refer to the MLEs of $M_1$ and $M_2$ when the scaled coalescent rates are fixed at their true values $C_1 = C_2 = C_3 = C_4 = 1$. Each box plot represents 1,000 repetitions of the simulation. The parameter values used to simulate the data in all cases were $C = 1.0$ and $M = M_1 = M_2 = 0.5$, indicated by the horizontal red lines.
Figure 14

(b)
unbiased, with increasing variance as the number of loci decreases. We also consider the case in which we are reasonably confident that $C_1 = C_2 = C_3 = C_4 = 1$ and wish to estimate $M_1$ and $M_2$ separately (Figure 14(a), bottom two plots). Again, our results suggest very good performance of our method in estimating the rates of gene flow in a three-species model, with unbiased estimates and decreasing variance as the number of loci increases.

The first simulation study indicates good performance of our method when gene trees are known without error. However, in the typical empirical setting, gene trees must first be estimated from observed sequence data (Figure 14(b),(c)). It is reasonable that compared to using directly simulated gene trees, the gene trees estimated from the sequence data produce less accurate estimates for both parameters. However, when the number of loci is large enough (more than 200), our method can still produce good estimates using the DNA sequence data. Also, the simulations with sequences of lengths 1,000bp led to better estimates than the simulations with 500bp, since more information is provided with longer DNA sequences and the gene tree estimates should be more accurate for longer genes. An interesting finding is that the simulations using sequence data always overestimate the gene flow rates when we assume $M_1 \neq M_2$. More specifically, when we assume that $C_1 = C_2 = C_3 = C_4 = 1$, we notice that $M_2$ is overestimated, while $M_1$ is only overestimated slightly when there are more than 100 loci. However, when the directly simulated gene trees are used to estimate the gene flow rates, neither $M_1$ and $M_2$ is overestimated consistently. One explanation for the overestimation is that if the speciation times are assumed known in our simulation studies, then incorrect gene trees can only be explained by increased migration rates. In fact, the incorrectness of the gene trees is generated due
to errors in reconstructing the gene trees from DNA sequences, especially for shorter DNA sequences. The increase in gene trees not matching the species tree could be incorrectly explained by either decreasing the time intervals between speciation times, or increasing gene flow rates. With constant speciation times assumed in our cases, the gene flow rates will be overestimated. It is not surprising that ancestral gene flow rates will be estimated with a larger bias, since it is more difficult to estimate deep coalescent events accurately.

3.4.5 Application to an empirical Afrotropical mosquito data set

Fontaine et al. (2015) reported pervasive autosomal gene introgression in several Afrotropical mosquito sibling species. In their study, the species branching order of seven Afrotropical mosquito sibling species was identified, and the times between speciation events were also estimated (Fontaine et al., 2015). We selected three of these species, Anopheles coluzzii (An. col), Anopheles gambiae (An. gam), and Anopheles arabiensis (An. ara), to serve as the species A, B, and C to test our model. We also selected an outgroup species, Anopheles christyi (An. chr), to root the gene trees. Based on Fontaine et al. (2015), the estimated species tree for the four species is (((An. col, An. gam), An. ara), An. chr). The speciation time between An. col and An. gam is 0.54 million years ago (Myr), and the speciation time between An. ara and the ancestor of An. col and An. gam is 1.85 Myr (Fontaine et al., 2015) This data set is a good test case for our methodology because previous work on this group provides reliable estimates of the species-level relationships and associated divergence times.
These data were analyzed by Wen et al. (2016) using methodology they developed for phylogenetic network inference and introgression detection. They sampled independent loci (genomic windows) from the MAF genome alignment of high-depth field samples. Since our method also assumes independence of the gene trees used as input, we downloaded the MAF genome alignment for the sampled loci used by Wen et al. (2016) from Dryad (doi: 10.5061/dryad.tn47c). In this data set, every pair of loci included were located at least 64 kb apart to fulfill the independence assumption, and the average length of the loci is about 3.4kb. We selected data from chromosome 2L, which includes a total of 669 independent alignments, from which gene trees were estimated in PAUP* using maximum likelihood. From the previous research (Fontaine et al., 2015), extensive gene introgression has been reported on chromosome 2L. Based on the speciation times given above, the frequencies of the coalescent histories were recorded. Under the assumption that all three selected *Anopheles gambiae* species have equal effective population sizes, the effective population size, the gene flow rate between *An. col* and *An. gam*, and the gene flow rate between *An. ara* and the ancestor of *An. col* and *An. gam* were estimated using maximum likelihood.

We searched for the MLEs for $\theta$, $M_1$, and $M_2$ in a two-step procedure. In the first step, we examined 60 equally-spaced values for $\theta$ ranging from 0.001 to 0.0594 and 60 equally-spaced values for both $\log(M_1)$, and $\log(M_2)$ ranging from -3 to 3, for a total of 360,000 values at which the likelihood was calculated. After finding that the likelihood was maximized along this grid at $M_1 = 0.1, M_2 = 19.9526$ and $\theta = 0.00675$, we used a finer grid that consisted of 200 equally-spaced values of $\log(M_2)$ ranging from -3 to 3, 200 equally-spaced values of $\theta$ ranging from 0.00396 to 0.01188, and four values of $M_1$ (0.01, 0.1, 1, and 10) (Figure 15).
Figure 15: Contour plots of the log likelihood for the coalescent rate $2/\theta$ and the gene flow rate $M_2$ at four different levels of gene flow rate $M_1$ assuming that $\theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta$. We analyzed 669 gene trees from chromosome 2L of the members of the Anopheles gambiae species complex. The estimated species tree topology and the speciation times were given by Fontaine et al. (2015). Four different levels of the gene flow rate between species A and B are shown in the panels: (a) $M_1 = 0.01$; (b) $M_1 = 0.1$; (c) $M_1 = 1$; (d) $M_1 = 10$. The MLEs are $\hat{\theta} = 0.00550$, $\hat{M}_1 = 0.1$, and $\hat{M}_2 = 7.413$, indicated by the white cross in panel (b), with log likelihood $l = -1,189.50$. 

76
The MLEs found in this manner were \( \hat{M}_1 = 0.1 \), \( \hat{M}_2 = 7.413 \), and \( \hat{\theta} = 0.00550 \). This result is highly consistent with the results of Fontaine et al. (2015) and Wen et al. (2016), in which they both conclude that there is substantial introgression between species An. ara and the ancestor of An. col and An. gam (Fontaine et al., 2015, Figure 1(c); Wen et al., 2016, Figure 5). Although introgression between An. col and An. gam is not tested in these two papers, our model suggests a lack of significant gene flow between these two species. This empirical study suggests that our model performs well in estimating the rates of coalescence and gene flow when the species tree, including speciation times, is well-estimated.

3.5 Discussion

3.5.1 Identifying the species tree from the gene tree topology distribution in the presence of gene flow

In the coalescent model for three species with no gene flow following speciation, the gene tree topology that matches the species tree will always have probability that is at least as large as that of the other two topologies, with equality only occurring when the time between the two speciation events is 0. When gene flow occurs between sister species, however, there are portions of the parameter space in which all three gene tree topologies have equal probability, even when the time interval between speciation events is not 0 (see Figure 11(c) and Figure 10(l)). Interestingly, we can characterize the portions of the parameter space for which this happens by the following: when there is substantial gene flow between sister lineages deeper in the tree, then substantial gene flow between sister species near the tips of the tree is required to identify the species tree topology from the gene tree topology distribution (compare Figure 11(c) to Figures 11(f), (i), and (l); compare Figure 10(l) to Figures
This finding is reasonable, because in the presence of a high rate of ancestral gene flow \((m_2)\) and the absence of gene flow elsewhere in the tree, the possible orders of coalescence among the three lineages will occur with approximately equal probability, and all three topologies will be equally likely, mimicking the case in which there is no gene flow and no time elapses between species events.

Our finding has important implications for species tree estimation. Several new methods for estimating species trees from large data sets have used the “rooted triples” method to build trees for subsets of the overall data set, with a second step in which the trees based on triplets are assembled into an overall species tree estimate (Ewing et al., 2008; DeGiorgio and Degnan, 2010; Liu et al., 2010; Poormohammadi et al., 2014). This method is expected to work well in the absence of gene flow, because the rooted triple with the highest probability under the coalescent model is known to be displayed on the species tree (Degnan et al., 2009). However, our result shows that in presence of gene flow (and more specifically, in the presence of gene flow more ancestrally in the rooted triple but not between the sister taxa near the tips), the rooted triple relationships cannot be accurately inferred given only topology frequencies. Adding stochastic variance due to the mutation process could result in misidentification of the correct rooted triple, biasing species tree inference methods based on this assumption when gene flow is present.

Finally, we recall our earlier result concerning identifiability of coalescent parameters along a fixed species tree in the presence of gene flow, noting again the contrast with results in the absence of gene flow. In particular, in the absence of gene flow, the probability distribution of gene tree topologies identifies both the species tree topology and associated speciation intervals (Allman et al., 2011a). Here we showed that,
even for a fixed species tree topology, many different coalescent parameter values may lead to the same distribution on gene tree topologies (Figure 12). This, too, has implications for species tree estimation, in that methods based only on the distribution of gene tree topologies cannot be used to infer species tree coalescent parameters. The distribution of coalescent histories, however, does appear to identify parameters, though we have not established this formally. This conjecture is supported by the positive performance of our method based on coalescent history distributions for both simulated and empirical data.

### 3.5.2 Applying the method to empirical data

Though there are many benefits in using the distribution of coalescent histories to estimate the coalescent parameters and the rates of gene flow, the application of this method has some limitations for empirical data. First, in order to classify the gene genealogies into different coalescent histories, a species tree with known speciation times is required. Though the species tree could first be estimated from the data, this would greatly increase the computational cost and possibly lead to biases in the ultimate estimates if variability in the estimated species tree is not properly accounted for in the subsequent estimate of the coalescent parameters. However, when a good estimate of the species tree and corresponding speciation times is given (as in our example data set of Afrotropical mosquitoes), our method can provide accurate estimates of the rates of gene flow and of the effective population sizes, parameters that are normally difficult to estimate.

Second, a fairly large number of loci are required to apply our method to an empirical data set. Our model provides the theoretical probability distribution of
coalescent histories, but these probabilities are not directly observable in practice; rather, they must be estimated from the observed frequencies of coalescent histories estimated from data. As the number of loci increases, the distribution of the observed coalescent histories will be closer to the theoretical distribution, in the absence of error in estimating the coalescent history. We showed the performance of our method in estimating parameters using different numbers of loci (Figure 14). For simulated gene trees, at least 100 loci are required to give reasonable estimates of the parameters. For simulated DNA sequences, at least 200 loci are required. For empirical data, we suggest using as many loci as possible to get good estimates of the effective population sizes and the gene flow rates. In our example for the Afrotropical mosquito data set, 669 independent gene trees across the alignment of chromosome 2L were constructed, and the parameters estimated from this large data set were very reasonable and consistent with previous work (Fontaine et al., 2015; Wen et al., 2016).

Finally, there is a computational burden incurred when working with this model due to the need for matrix exponentiation and numerical integration. Computations are feasible when the values of the coalescent parameter $\theta_x$ and the speciation time $\tau_y$ are in a reasonable range. The reference species tree we used in this paper follows the parameters used in the research of Zhu and Yang (2012), with $\theta_A = \theta_B = \theta_C = \theta_{AB} = 0.005$, $\tau_1 = 0.004$, and $\tau_2 = 0.006$. We suggest keeping the ratio of $\tau/\theta$ less than 10 to avoid any numerical issues. These issues can likely be overcome by implementing more sophisticated methods for calculating the matrix exponential and the numerical integrals.

We note that the IMa method of Hey (2010) will be more generally appropriate for inference of speciation history in the presence of gene flow for several reasons. First, it
takes as input the sequence data directly via utilization of an MCMC-based Bayesian inferential framework. Second, it does not require knowledge of the timing of speciation events. Similar methods considering mutational patterns are also introduced in Gutenkunst et al. (2009), Lohse et al. (2011), Lohse and Frantz (2014), and Kamm et al. (2015). However, our intention in presenting the application to the mosquito data set was two-fold: (1) to provide an example of the use of the calculations above in a practical setting, with the goal of aiding in understanding the significance of the results; and (2) to provide a complementary method of parameter estimation that could be used alongside sequence-based methods as validation of the appropriateness of the underlying models and methods.

3.5.3 Future directions

Our model is constructed for three species with gene flow between both pairs of sister species. The model can be extended to an arbitrary number of species with gene flow between all sister-species pairs by constructing larger instantaneous rate matrices for each time period and increasing the dimension of integrals. In this case, there would also be more time intervals along the species tree that would need to be considered. For example, the largest instantaneous rate matrix for a four-species bifurcating tree will be $29 \times 29$, and it would require a three-dimensional integral. While it is not difficult to list all the density functions and the marginal probability functions as we have done here, the computational cost of calculating these quantities grows rapidly. A spectral decomposition method similar to that used in Andersen et al. (2014) could be an effective way to overcome the computational burden. In their paper, the spectral decomposition method was used to model a scenario in which
an ancestral population splits into P subpopulations at some time $T_A$ in the past (Andersen et al., 2014). A similar method of dividing a rate matrix into several submatrices could be helpful in implementing our model for an arbitrary number of species.

Another extension of our model is to add more sequences for each species. As in the case of adding more species discussed above, adding more lineages will also increase the dimension of the integrals as well as the size of the instantaneous rate matrices. Again, the main issue is improving the computational method so that the numerical probabilities of each coalescent history can be calculated efficiently.

A further extension of our model would be a model in which gene flow can occur globally throughout the phylogeny, rather than simply between sister species. This would increase biological realism, because though it is possible that most gene flow happens between closely related species, it is also possible that gene flow exists in more distantly related species. As shown in Figure 3.1, our model assumes no gene flow between species A and C, and species B and C from the present to time $\tau_1$. If gene flow existed to some extent between species A and C, and species B and C, additional coalescent histories would be possible, and the symmetric probabilities of topologies G2 and G3 may be affected by the induced gene flow. To implement a model with more widespread gene flow, we need to introduce a new instantaneous rate matrix to describe the coalescent and the gene flow process in the time interval from the present to time $\tau_1$, and then carry out calculations along the lines of what we did here for that new matrix. While the methods are straightforward, computational challenges are the main limitation of this approach.
3.5.4 Conclusions

This article presents a method for computing the coalescent history distribution under the coalescent model for three species that allows gene flow between both pairs of sister populations. The ability to compute coalescent history distributions for species trees with various effective population sizes as well as various gene flow rates leads to a better understanding of evolutionary relationships among closely related populations or species. The application of the coalescent history distributions in simulation studies as well as to empirical data sets allows us to infer species tree parameters, such as the ancestral effective population sizes and the gene flow rates, using maximum likelihood. This study also demonstrates that for a fixed species tree topology, many different coalescent parameter values may lead to the same distribution on gene tree topologies, while the distribution of the coalescent histories is distinct for different choices of parameters. These findings have implications for the development of species tree estimation methods in the presence of gene flow. Future work is needed to formally establish conditions for identifiability of the species tree from the coalescent history distribution, as well as to extend the coalescent model with gene flow to an arbitrary number of species with more than one sampled gene per species.
Figure 16: The probability distribution of the gene tree histories under species trees with scaled speciation times $T_1 = 1.6$, $T_2 = 2.4$ ($\tau_1 = 0.004$, $\tau_2 = 0.006$, $\theta_0 = 0.005$) for panels (c) and (d), and $T_1 = 2$, $T_2 = 4$ ($\tau_1 = 0.01$, $\tau_2 = 0.02$, $\theta_0 = 0.01$) for panels (a), (b), (e) and (f). Each gene tree history is denoted by a different color as shown in the figure. The probability of topology G1 is shown by the height of the column labeled G1; the height of the column labeled G2/3 shows the equal probability of the topologies G2 and G3. Thus, $P(G1) + 2P(G2/3) = 1$. The two sets of the scaled coalescent rates are $C_1 = C_2 = C_3 = C_4 = 1$ (scaled by $\theta_0 = 0.005$) in panels (a), (c), and (e), and $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2$ (scaled by $\theta_0 = 0.005$) in panels (b), (d) and (f). Each panel contains four cases of different gene flow rates: 1, no gene flow ($M_1 = M_2 = 0$); 2, no gene flow between species A and B ($M_1 = 0$, $M_2 = 0.5$ for panels (a) and (b); $M_1 = 0$, $M_2 = 2$ for panels (c) - (f)); 3, no gene flow between species C and the ancient species AB ($M_1 = 0.5$, $M_2 = 0$ for panels (a) and (b); $M_1 = 2$, $M_2 = 0$ for panels (c) - (f)); 4, equal rates of gene flow in both sister species ($M_1 = M_2 = 0.5$ for panels (a) and (b); $M_1 = M_2 = 2$ for panels (c) - (f)). Note that all other parameters (coalescent rates and gene flow rates) are the same in Figure 9 (a) and Figure 16 (b).
Figure 17: Probability distributions of the gene tree histories for the model of three species with gene flow between sister species. The probabilities of each gene tree history (y-axis) were plotted against the gene flow rate between species C and the ancient species AB $M_2$ (x-axis; shown on a log scale). The four sets of parameter values are $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 1.6$, $T_2 = 2.4$, $\theta_0 = 0.005$, $M_1 = 0.001$ for panels (a) - (c); $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 1.6$, $T_2 = 2.4$, $\theta_0 = 0.005$, $M_1 = 0.5$ for panels (d) - (f); $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 1.6$, $T_2 = 2.4$, $\theta_0 = 0.01$, $M_1 = 2$ for panels (g) - (i); and $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 1.6$, $T_2 = 2.4$, $\theta_0 = 0.005$, $M_1 = 20$ for panels (j) - (l).
Figure 17

(a) 

(b) 

(c) 

(d) 

(e) 

(f) 

(g) 

(h) 

(i) 

(j) 

(k) 

(l) 

Migration rate between AB and C
Figure 18: Probability distributions of the gene tree histories for the model of three species with gene flow between sister species. The probabilities of each gene tree history (y-axis) were plotted against the gene flow rate between species C and the ancient species AB $M_2$ (x-axis; shown on a log scale). The four sets of parameter values are $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.005$, $M_1 = 0.001$ for panels (a) - (c); $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.005$, $M_1 = 0.5$ for panels (d) - (f); $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.01$, $M_1 = 2$ for panels (g) - (i); and $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.005$, $M_1 = 20$ for panels (j) - (l).
Figure 18

(a) 
(b) 
(c) 

(d) 
(e) 
(f) 

(g) 
(h) 
(i) 

(j) 
(k) 
(l)
Figure 19: Probability distributions of the gene tree histories for the model of three species with gene flow between sister species. The probabilities of each gene tree history (y-axis) were plotted against the gene flow rate between species C and the ancient species AB $M_2$ (x-axis; shown on a log scale). The four sets of parameter values are $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2, T_1 = 1.6, T_2 = 2.4, \theta_0 = 0.005$, $M_1 = 0.001$ for panels (a) - (c); $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2, T_1 = 1.6, T_2 = 2.4, \theta_0 = 0.005, M_1 = 0.5$ for panels (d) - (f); $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2, T_1 = 1.6, T_2 = 2.4, \theta_0 = 0.01, M_1 = 2$ for panels (g) - (i); and $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2, T_1 = 1.6, T_2 = 2.4, \theta_0 = 0.005, M_1 = 20$ for panels (j) - (l).
Figure 19

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)
Figure 20: Probability distributions of the gene tree histories for the model of three species with gene flow between sister species. The probabilities of each gene tree history (y-axis) were plotted against the gene flow rate between species C and the ancient species AB $M_2$ (x-axis; shown on a log scale). The four sets of parameter values are $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.005$, $M_1 = 0.001$ for panels (a) - (c); $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.005$, $M_1 = 0.5$ for panels (d) - (f); $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.01$, $M_1 = 2$ for panels (g) - (i); and $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.005$, $M_1 = 20$ for panels (j) - (l).
Figure 20

(a) 

(b) 

(c) 

(d) 

(e) 

(f) 

(g) 

(h) 

(i) 

(j) 

(k) 

(l)
Figure 21: Probability distributions of the gene tree histories for the model of three species with gene flow between sister species. The probabilities of each gene tree history (y-axis) were plotted against the gene flow rate between species A and species B $M_1$ (x-axis; shown on a log scale). The four sets of parameter values are $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 1.6$, $T_2 = 2.4$, $\theta_0 = 0.005$, $M_2 = 0.001$ for panels (a) - (c); $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 1.6$, $T_2 = 2.4$, $\theta_0 = 0.005$, $M_2 = 0.5$ for panels (d) - (f); $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 1.6$, $T_2 = 2.4$, $\theta_0 = 0.005$, $M_2 = 2$ for panels (g) - (i); and $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 1.6$, $T_2 = 2.4$, $\theta_0 = 0.005$, $M_2 = 20$ for panels (j) - (l).
Figure 21
(a) (b) (c)
(d) (e) (f)
(g) (h) (i)
(j) (k) (l)
Figure 22: Probability distributions of the gene tree histories for the model of three species with gene flow between sister species. The probabilities of each gene tree history (y-axis) were plotted against the gene flow rate between species A and species B $M_1$ (x-axis; shown on a log scale). The four sets of parameter values are $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.005$, $M_2 = 0.001$ for panels (a) - (c); $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.005$, $M_2 = 0.5$ for panels (d) - (f); $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.005$, $M_2 = 2$ for panels (g) - (i); and $C_1 = C_2 = C_3 = C_4 = 1$, $T_1 = 2$, $T_2 = 4$, $\theta_0 = 0.005$, $M_2 = 20$ for panels (j) - (l).
Figure 22

(a) 

(b) 

(c) 

(d) 

(e) 

(f) 

(g) 

(h) 

(i) 

(j) 

(k) 

(l)
Figure 23: Probability distributions of the gene tree histories for the model of three species with gene flow between sister species. The probabilities of each gene tree history (y-axis) were plotted against the gene flow rate between species A and species B $M_1$ (x-axis; shown on a log scale). The four sets of parameter values are $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2, T_1 = 1.6, T_2 = 2.4, \theta_0 = 0.005, M_2 = 0.001$ for panels (a) - (c); $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2, T_1 = 1.6, T_2 = 2.4, \theta_0 = 0.005, M_2 = 0.5$ for panels (d) - (f); $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2, T_1 = 1.6, T_2 = 2.4, \theta_0 = 0.005, M_2 = 2$ for panels (g) - (i); and $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2, T_1 = 1.6, T_2 = 2.4, \theta_0 = 0.005, M_2 = 20$ for panels (j) - (l).
Figure 23

(a) 

(b) 

(c) 

(d) 

(e) 

(f) 

(g) 

(h) 

(i) 

(j) 

(k) 

(l)
Figure 24: Probability distributions of the gene tree histories for the model of three species with gene flow between sister species. The probabilities of each gene tree history (y-axis) were plotted against the gene flow rate between species A and species B $M_1$ (x-axis; shown on a log scale). The four sets of parameter values are $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2, T_1 = 2, T_2 = 4, \theta_0 = 0.005, M_2 = 0.001$ for panels (a) - (c); $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2, T_1 = 2, T_2 = 4, \theta_0 = 0.005, M_2 = 0.5$ for panels (d) - (f); $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2, T_1 = 2, T_2 = 4, \theta_0 = 0.005, M_2 = 2$ for panels (g) - (i); and $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2, T_1 = 2, T_2 = 4, \theta_0 = 0.005, M_2 = 20$ for panels (j) - (l).
Chapter 4: Rooting phylogenetic trees under the coalescent model using site pattern probabilities

Phylogenetic tree inference is a fundamental tool to estimate ancestor-descendant relationships among different species. In phylogenetic studies, identification of the root - the most recent common ancestor of all sampled organisms - is essential for complete understanding of the evolutionary relationships. Rooted trees benefit most downstream application of phylogenies such as species classification or study of adaptation. Often, trees can be rooted by using outgroups, which are species that are known to be more distantly related to the sampled organisms than any other species in the phylogeny. However, outgroups are not always available in evolutionary research. In this study, we develop a new method for rooting species tree under the coalescent model, by developing a series of hypothesis tests for rooting quartet phylogenies using site pattern probabilities. The power of this method is examined by simulation studies and by application to an empirical North American rattlesnake data set. Our study establishes a computationally practical method for rooting species trees that is more efficient than traditional methods. The method will benefit numerous evolutionary studies that require rooting a phylogenetic tree without having to specify outgroups.
4.1 Introduction

Phylogenetic tree inference is a fundamental framework in which to estimate the ancestor-descendant relationships among different species. Currently, the amount of DNA sequence data is increasing dramatically, and more accurate and efficient methods are required to estimate phylogenetic trees using these data. Evolutionary relationships can be analyzed at two distinct levels (gene trees and species trees), and it is not necessary for the gene trees and species trees to agree with one another (Pamilo and Nei, 1988; Takahata, 1989; Hein, 1993; Maddison, 1997; Kubatko, 2009; Bayzid and Warnow, 2012). Incomplete lineage sorting (ILS) is considered to be one of the major factors that causes disagreement between species trees and gene trees, and thus ILS has a critical effect on estimation of the species tree using large multilocus data sets (Kingman, 1982a; Tajima, 1983; Tavaré, 1984; Takahata and Nei, 1985; Pamilo and Nei, 1988; Rosenberg, 2002; Rannala and Yang, 2003; Degnan and Salter, 2005).

In many species tree inference approaches, gene trees are estimated first and are assumed as known in the following analysis (Mirarab et al., 2014; Mirarab and Warnow, 2015; Maddison and Knowles, 2006; Than and Nakhleh, 2009; Liu et al., 2010; Fan and Kubatko, 2011; Liu et al., 2009; Wu, 2012; Liu and Yu, 2011). However, it is highly possible that the estimated gene trees maybe either biased or not fully informative, because they are often based on short sequences with few variable sites (Brower et al., 1996). As a result, the gene tree estimation errors may potentially become a severe issue in species tree inference. Some coalescent inference methods, such as ASTRAL, does not directly infer the root of the estimated species phylogeny (Mirarab et al., 2014; Mirarab and Warnow, 2015). Still other coalescent inference
Figure 25: Example species trees. All possible splits of a four-leaf unrooted species tree are shown in the center of the figure. Five possible root positions are labelled on the unrooted tree, with arrows pointing to the five possible rooted species trees with four leaves.
methods (MP-EST, NJst) require rooted gene trees as the input in order to estimate a rooted species tree (Liu et al., 2010; Liu and Yu, 2011). However, ancestor (rooting) identification is essential for complete understanding of the evolutionary relationships. Rooted trees benefit most downstream applications of phylogenies, such as species classification and comparative biology. In many cases, trees can be rooted using outgroups, which are known species that are more distantly related to the sampled organisms than any other species in the phylogeny. However, outgroups are not always available in evolutionary research. For instance, in numerous unresolved evolutionary questions such as animal evolution (Aguinaldo et al., 1997; Philippe et al., 2011), placental mammal evolution (Waddell et al., 1999; Madsen et al., 2001; Murphy et al., 2001; Scally et al., 2001), prokaryotic evolution (Cox et al., 2008; Lake, 2009), and even the beginnings of life (Lake, 2009; Ragan et al., 2009), it is difficult to specify outgroups. Thus, rooting methods in the absence of outgroups are often necessary for phylogenetic inference.

In our study, we develop a new method for rooting species tree under the coalescent model, by developing a series of hypothesis tests for rooting quartet phylogenies using site pattern probabilities. More specifically, the site pattern probabilities of every four-taxon quartet are used to construct rooted species trees based on an unrooted species tree topology. Our study establishes a computationally practical method of rooting species trees in the absence of an outgroup. Since a rooted species tree will provide more information about evolutionary relationships, the new method will benefit numerous evolutionary studies that require rooting a phylogenetic tree without having to specify outgroups.
4.2 Methods

The coalescent process (Kingman, 1982a; Takahata and Nei, 1985; Pamilo and Nei, 1988) is a retrospective model of population genetics that is commonly used to model incomplete lineage sorting (ILS). Based on tracing the evolutionary history of sampled genes by considering the time from the present back to their most recent common ancestor (Kingman, 2000), the coalescent model is used as the basis for different methods to estimate species trees (e.g. (Kubatko et al., 2009; Than et al., 2007; Liu and Pearl, 2007; Liu et al., 2010; Heled and Drummond, 2010); reviewed in Edwards (2009) ). Under the coalescent model, our method uses relationships among the expected site pattern probability to develop a method to root phylogenetic trees. Considering each column of a DNA alignment that corresponds to an evolutionary process whereby these nucleotides evolved from a single ancestral nucleotide as a site, when a set of sites under the coalescent model are assumed to freely recombine, they are defined as coalescent independent sites.

4.2.1 Method for Rooting Phylogenetic Trees by Site Pattern Probabilities

In a four-taxon species tree, there are $4^4 = 256$ possible site patterns. Let $p_{i_Ai_Bi_Ci_D}, (i_a \in \{A, C, G, T\}, a = A, B, C, D)$ represent the probabilities of each site pattern $i_Ai_Bi_Ci_D$, where $i_a$ refers to the nucleotide at tip $a$ of the four-taxon species tree. Any site pattern probability of a rooted four-taxon species tree under the molecular clock assumption can be classified into one of 15 categories:

$$P_{xxxx}, P_{xxxy}, P_{xxyx}, P_{xyxx}, P_{yxxx}, P_{xyty}, P_{xyxy}, P_{xxyy}, P_{xyyx}, P_{xyyy},$$
where \( w, x, y, \) and \( z \) denote different nucleotide states. To explore the rooting position for an unrooted four-taxon tree, which can then be used to infer the root position on a larger phylogenetic tree, we develop a series of hypothesis tests, based on expected site pattern probabilities (Table 1).

Table 1: Relationships expected among site pattern probabilities for various root positions.

<table>
<thead>
<tr>
<th>Rooting position</th>
<th>Decision rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( p_{yxxx} &gt; p_{xyxx}, p_{xxx} = p_{xxy} )</td>
</tr>
<tr>
<td>2</td>
<td>( p_{xyxx} &gt; p_{yxxx}, p_{xxx} = p_{xxy} )</td>
</tr>
<tr>
<td>3</td>
<td>( p_{xxy} &gt; p_{xyxx}, p_{yxxx} = p_{xxx} )</td>
</tr>
<tr>
<td>4</td>
<td>( p_{xxx} &gt; p_{xyxx}, p_{xyxx} = p_{xxx} )</td>
</tr>
<tr>
<td>5</td>
<td>( p_{xyxx} = p_{yxxx}, p_{xxx} = p_{xxy} )</td>
</tr>
</tbody>
</table>

These hypothesis tests are derived from the equivalence of site pattern probabilities in a four-taxon phylogenetic tree. For instance, if the rooting position is 1 (Figure 25), it is clear that species C and D have equal probabilities of mutating under the molecular clock assumption. Therefore, \( p_{xxx} = p_{xxy} \). On the other hand, species A can be considered as an outgroup in these four species, and the site pattern \( yxxx \) is more likely than \( xyxx \), thus it is easy to see that \( p_{yxxx} > p_{xyxx} \). Similarly, we can write expected relationships for the other four root positions (Table 1, Figure 25). Note
that $p_{yzxz} = p_{zyxx}$, and $p_{yxzx} = p_{gyxx}$, could also be used, but our preliminary results suggested that the values of $p_{zyxx}, p_{yxzx}, p_{xxxy}$ and $p_{xxyx}$ are larger than $p_{zyxz}, p_{xyzx}, p_{yxxz}$, and $p_{yxzx}$, thereby giving better performance when estimated from empirical data.

Note that the analytical derivation of the site pattern probabilities arising from the coalescent model under the JC69 model is given by Chifman and Kubatko (2015). It is not surprising that under the JC69 model, many site pattern probabilities are identical due to the assumption of equal base frequencies and identical nucleotide substitution rates. Indeed, site pattern probabilities within each category described above are identical under the JC69 model. Therefore, based on the precise formulas for the site pattern probabilities derived by Chifman and Kubatko (2015), the relationships in Table 1 can be mathematically proved under the JC69 model. Analytical proof is not given for other nucleotide substitution models due to increasing complexity in computing caused by unequal base frequencies and varying nucleotide substitution rates. However, with the clock assumption, it is still reasonable to apply the method under other nucleotide substitution models, because the probabilities of having mutations are identical for sister species and are always proportional to branch length under any nucleotide substitution model. The performance of our rooting method under varying nucleotide substitution models will be tested using simulation studies.

### 4.2.2 Formal hypothesis tests

To determine the root position, we first set up two distinct hypothesis tests:

Test 1: $H_0 : p_{yxxx} = p_{zyxx}$ vs. $H_1 : p_{yxxx} \neq p_{zyxx}$.
and

Test 2: $H_0: p_{xxxy} = p_{xxxy}$ vs. $H_1: p_{xxxy} \neq p_{xxxy}$.

Note that there are 12 possible site pattern probabilities within each category of $yxxx$, $xyxx$, $xxyx$, or $xxxy$, and the average site pattern probabilities of each category can be selected to form the test statistics for our rooting method. Let $\mathbf{X} = [X_1, X_2, ..., X_{256}]$ denote the vector of the counts for each site pattern, and $\mathbf{p} = [p_1, p_2, ..., p_{256}]$ denote the vector of the 256 site pattern probabilities. Then $\mathbf{X} \sim \text{Multinomial}(M, \mathbf{p})$, where $M$ is the number of coalescent independent sites. Under the assumption of a multinomial distribution, we can compute the mean and variance of each count and the covariance between them. First, note that

\[
E(X_s) = Mp_s \quad s = 1, 2, ..., 256,
\]

\[
\text{Var}(X_s) = Mp_s(1 - p_s) \quad s = 1, 2, ..., 256,
\]

\[
\text{cov}(X_s, X_t) = -Mp_sp_t \quad s = 1, 2, ..., 256, t = 1, 2, ..., 256, s \neq t.
\]

Also, to simplify the expression, let

\[
p_a = \frac{1}{12} \sum_{i,j \in \{A,C,G,T\}, i \neq j} p_{jiii}, \quad \hat{p}_a = \frac{1}{12} \sum_{i,j \in \{A,C,G,T\}, i \neq j} \hat{p}_{jiii},
\]

\[
p_b = \frac{1}{12} \sum_{i,j \in \{A,C,G,T\}, i \neq j} p_{ijii}, \quad \hat{p}_b = \frac{1}{12} \sum_{i,j \in \{A,C,G,T\}, i \neq j} \hat{p}_{ijii},
\]

\[
p_c = \frac{1}{12} \sum_{i,j \in \{A,C,G,T\}, i \neq j} p_{iiij}, \quad \hat{p}_c = \frac{1}{12} \sum_{i,j \in \{A,C,G,T\}, i \neq j} \hat{p}_{iiij},
\]

\[
p_d = \frac{1}{12} \sum_{i,j \in \{A,C,G,T\}, i \neq j} p_{iiij}, \quad \hat{p}_d = \frac{1}{12} \sum_{i,j \in \{A,C,G,T\}, i \neq j} \hat{p}_{iiij},
\]

where $\hat{p}_{jiii}$ denotes the estimated probability of site pattern $jiii$ from the observed data, for example.
Thus, we have:

$$E(\hat{p}_a - \hat{p}_b) = \frac{1}{12} \left( \sum_{i,j \in \{A,C,G,T\}, i \neq j} \hat{p}_{jii} - \sum_{i,j \in \{A,C,G,T\}, i \neq j} \hat{p}_{jii} \right) = p_a - p_b$$  \hspace{1cm} (4.1)$$

$$Var(\hat{p}_a - \hat{p}_b) = \frac{1}{12M} \left[ p_a(1 - p_a) + p_b(1 - p_b) - 2p_ap_b \right]$$  \hspace{1cm} (4.2)$$

$$E(\hat{p}_c - \hat{p}_d) = \frac{1}{12} \left( \sum_{i,j \in \{A,C,G,T\}, i \neq j} \hat{p}_{iiij} - \sum_{i,j \in \{A,C,G,T\}, i \neq j} \hat{p}_{iiij} \right) = p_c - p_d$$  \hspace{1cm} (4.3)$$

$$Var(\hat{p}_c - \hat{p}_d) = \frac{1}{12M} \left[ p_c(1 - p_c) + p_d(1 - p_d) - 2p_cp_d \right]$$  \hspace{1cm} (4.4)$$

Now, using Equations (4.1) - (4.4), substituting the estimated site pattern probabilities into Equations (4.2) and (4.4), we can compute test statistics for both hypothesis tests:

$$Z_1 = \frac{\hat{p}_a - \hat{p}_b}{\sqrt{\frac{1}{12M} \left[ \hat{p}_a(1 - \hat{p}_a) + \hat{p}_b(1 - \hat{p}_b) - 2\hat{p}_a\hat{p}_b \right]}}$$  \hspace{1cm} (4.5)$$

$$Z_2 = \frac{\hat{p}_c - \hat{p}_d}{\sqrt{\frac{1}{12M} \left[ \hat{p}_c(1 - \hat{p}_c) + \hat{p}_d(1 - \hat{p}_d) - 2\hat{p}_c\hat{p}_d \right]}}$$  \hspace{1cm} (4.6)$$

Under the null hypothesis in Test 1 that \( p_{yxxx} = p_{xxyxx} \), \( Z_1 \sim N(0,1) \) when \( M \) is large. Similarly, \( Z_2 \sim N(0,1) \) under the null hypothesis in Test 2 that \( p_{xxyx} = p_{xxy} \) when \( M \) is large. Therefore, our rooting method can be applied by checking the test results and values of \( Z_1 \) and \( Z_2 \). More specifically, for example, if we reject Test 1, accept Test 2, and \( Z_1 > 0 \), we can conclude that the root position is \( \hat{1} \). Similarly, the other test results and their conclusions are summarized in Table 2. Note that significance levels for the two tests, \( \alpha_1 \) and \( \alpha_2 \), must be selected. In our study, we choose the significance levels \( \alpha_1 = \alpha_2 = 0.025 \). The significance levels can be

110
adjusted for different studies. The performance of the rooting method are evaluated by simulation studies, as described below.

Table 2: Test results and conclusions for rooting a four-taxon phylogenetic tree.

<table>
<thead>
<tr>
<th>Test 1</th>
<th>Test 2</th>
<th>$Z_1$</th>
<th>$Z_2$</th>
<th>Inferred rooting position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject</td>
<td>Accept</td>
<td>$&gt; 0$</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>Reject</td>
<td>Accept</td>
<td>$&lt; 0$</td>
<td>NA</td>
<td>2</td>
</tr>
<tr>
<td>Accept</td>
<td>Reject</td>
<td>NA</td>
<td>$&gt; 0$</td>
<td>3</td>
</tr>
<tr>
<td>Accept</td>
<td>Reject</td>
<td>NA</td>
<td>$&lt; 0$</td>
<td>4</td>
</tr>
<tr>
<td>Accept</td>
<td>Accept</td>
<td>NA</td>
<td>NA</td>
<td>5</td>
</tr>
<tr>
<td>Reject</td>
<td>Reject</td>
<td>NA</td>
<td>NA</td>
<td>No conclusion</td>
</tr>
</tbody>
</table>

4.2.3 Simulation Studies

Three sets of simulation studies were used to examine the performance of our method to root the species quartets. All simulation studies include DNA sequence data simulated from four-taxon species trees. More specifically, different numbers of gene trees are generated from the species trees with COAL (Degnan and Salter, 2005), then coalescent independent sites or multi-locus DNA sequences are simulated by using Seq-Gen (Rambaut and Grassly, 1997). The simulation process is repeated 100 times to generate 100 independent data sets, the rooting method is applied to each data set, and the power (proportion of the 100 data sets for which the correct conclusion is made) for each simulation setting is recorded.

The first set of simulation studies is designed to assess the performance of our test for coalescent independent sites with a constant evolutionary rate across the sites. Two groups of species trees with “long” and “short” branch lengths are used to
simulate the data. Each group contains five species trees that have the same unrooted topology, but different rooting positions. For the “long branch lengths” group, the five species trees are 

\((A : 3.0, (B : 2.0, (C : 1.0, D : 1.0) : 1.0)), (B : 3.0, (A : 2.0, (C : 1.0, D : 1.0) : 1.0)), (C : 3.0, (D : 2.0, (A : 1.0, B : 1.0) : 1.0)), (D : 3.0, (C : 2.0, (A : 1.0, B : 1.0) : 1.0)),\)

and 

\((A : 0.8, B : 0.8) : 2.2, (C : 1.2, D : 1.2) : 1.8).\)

The five species trees in the “short branch lengths” group have the same topologies as in the “long branch lengths” group, but all branch lengths are scaled by 0.5 (all branch lengths in our study are measured in coalescent units). A varying number of gene trees (5,000, 10,000, 20,000, 100,000) are simulated from each species tree. The gene trees are then used to simulate coalescent independent sites under the JC69, HKY85, and GTR+I+\(\Gamma\) models. For each parameter setting, 100 replications are simulated to estimate the root position, and the proportion for which the correct conclusion is reached is recorded as the power of the study.

The second set of simulation studies focuses on multi-loci DNA sequence data instead of coalescent independent sites. In the first set of simulation studies, we simulate a number of gene trees, and only one site is simulated under each gene tree as a coalescent independent site. However, we also wish to explore the performance of our method for multi-locus data. The simulation studies have similar parameter settings to the first set of simulations, but instead of a single site, a DNA sequence of 100 base pairs is simulated from each gene tree by Seq-Gen. The number of gene trees is adjusted to (50, 100, 200, 1000) to keep the total number of sites identical to that used in the first set of simulations.

The last set of simulation studies uses coalescent independent sites. It has parameter settings similar to those in the first set of simulations, but will assess the
effect of rate variation across lineages. More specifically, after the gene trees are simulated from the given species trees, an extra step that draws varying evolutionary rates from a lognormal distribution with mean = 1 is added to assign different evolutionary rates across the branches of the tree. Three lognormal distributions are applied in our simulation studies with the parameter settings: \( m = -0.005, s = 0.1; m = -0.125, s = 0.5; m = -0.5, s = 1.0 \). The performance of our method for rooting the phylogenetic trees from coalescent independent sites with varying evolutionary rates across the tree is tested in this set of simulation studies.

### 4.2.4 Application to Larger Species Trees

To examine the performance of our rooting method for larger taxon samples, we assume that the unrooted tree has been previously estimated. In our example, we estimate the species tree using SVDQuartets, a full-data coalescent-based method based on site pattern probabilities, and we label each branch with a particular code (Figure 26a). Our method works by randomly selecting a subset of four species from the \( n \) species under study, and determining the root position, as shown in Figure 25. This is repeated many times, for many randomly selected quartets. If the number of taxa is not too large, all quartets can be considered; otherwise, a random sample can be taken. Note that there are multiple correlated hypothesis tests for a species tree with more than 4 taxa. To handle the issue of multiple tests, we use the Bonferroni correction. When an overall \( \alpha \)-level test for an \( n \)-taxon species tree is desired, we use \( \alpha / \binom{n}{4} \) as the critical value in the tests, when all quartets are sampled.

To determine the root of a given species tree with more than 4 taxa after the selected quartets have been evaluated, we develop a method to combine the results
Figure 26: Example of scoring potential root positions on larger phylogenies. (a) A six-leaf unrooted species tree. The branches are coded from $a$ to $i$, and the internal branches are highlighted in red; (b) An example species quartet ABCF. If the root is determined to be on branch $a$, $f$ or $h$, the corresponding branch on the tree in (a) will get score 1. If the root is determined to be on branch $b + g$ (or $c + l$), branches $b$ and $g$, (or $c$ and $l$) in (a) will each get score 0.5. (c) An example species quartet AECF. If the root is determined to be on branch $a$, $c$ or $d$, the corresponding branch in (a) will get score 1. Otherwise, if the root is determined to be on branch $e + g$ (or $h + l$), branches $e$ and $g$ (or $h$ and $l$) will each get score 0.5.
from the individual quartet tests. This method assigns a weighted score for each branch based on the results of the analysis of the individual quartets. Suppose a particular species quartet is composed of five branches (Figure 26a, b), where any branch contains one or more coded branches as shown in Figure 26a. Denote the number of the coded branches within the five branches as $n_1$, $n_2$, $n_3$, and $n_4$, $n_5$, respectively. Once a branch $n_i$ ($i = 1, 2, 3, 4, 5$) is determined to contain the root, any coded branch within the determined branch has score $\frac{1}{n_i}$, while other branches have score 0. Two examples are shown in Figures 26b and 26c. The branch with the highest summed scores over all quartets evaluated will be selected as the location of the root.

4.3 Results

4.3.1 Accuracy of the method for rooting phylogenetic trees

The power of the rooting method in the three simulation studies is shown in Figure 27. In 100 simulations, the proportion of the data sets for which the correct rooting position is selected is summarized. The panels in the first column of Figure 27 (panels (a), (d), (g), (j), and (m)) represent the power for detecting the correct root positions for the simulation studies with coalescent independent sites with a constant evolutionary rate across the sites. The panels in the middle column (panels (b), (e), (h), (k), and (n)) show the power for rooting phylogenetic trees in the second simulation set, where multi-loci DNA sequence data is simulated with a constant evolutionary rate. The last column (panels (c), (f), (i), (l), and (o)) shows the results of the third set of simulation studies, where coalescent independent sites are simulated with varying evolutionary rates. Clearly, the simulation conditions that
strictly follow the assumptions (free recombination and constant evolutionary rate) of the rooting method have very high power. When the assumption of free recombination is violated (e.g., for the multi-loci DNA sequence data in column 2), the tests have a slightly lower accuracy when the number of sites is small. If only the assumption of constant evolutionary rate is violated, our method still gives high power to detect the root position. In fact, when varying evolutionary rates are drawn from a lognormal distribution with mean = 1, the power of our method is not different than the first set of simulation studies. Overall, it is safe to conclude that the new rooting method has high accuracy for rooting a four-leaf unrooted species tree. Notably, when the sample size is increased to about 10,000 bp, the accuracy is over 90% even for multi-loci DNA sequence data.

In our simulation studies, DNA sequences data are simulated under three different nucleotide substitution models: JC69, HKY85, and GTR+I+Γ (labeled by black, red, and green in Figure 27). Though the hypothesis tests for the rooting method are derived from the JC69 model, as described in the Methods section, the results of the simulation studies suggests that the rooting method can be applied to more general nucleotide substitution models. Over all of the conditions we tested, there was no systematic difference between the results for the JC69 model and for the other two models. Furthermore, the performance of our method in rooting the phylogenetic trees depends primarily on the sample size. More specifically, species trees with more coalescent independent sites or longer DNA sequences sampled can be rooted more accurately. As shown in Figure 27, the solid lines denote the results for the species trees with “longer” branch lengths, and the dashed lines show the results for the species trees with “shorter” branch lengths. In general, the power for species trees
Figure 27: Accuracy of the rooting method. In each panel, the x-axis denotes the data size (kb), and the y-axis shows the proportion of the data sets for which the correct rooting position is selected in a total of 100 simulations. From left to right, each column represents one set of simulation studies as described in the Methods section. (a), (d), (g), (j), (m): The first set of simulation studies (coalescent independent sites, without rate variation across lineages); (b), (e), (h), (k), (n): The second set of simulation studies (multi-loci DNA, without rate variation across lineages); (c), (f), (i), (l), (o): The third set of simulation studies (coalescent independent sites, with rate variation across lineages). From top to bottom, each row represents one possible split of a four-leaf unrooted species tree denoted by 1 - 5 as in Figure 25. (a) - (c): Root position at 1, (d) - (f): Root position at 2, (g) - (i): Root position at 3, (j) - (l): Root position at 4, (m) - (o): Root position at 5. Solid lines in each panel represent the species trees in the “Long branch lengths” group, while the species trees in the “Short branch lengths” group are denoted by dashed lines. In the simulation studies, DNA sequences data is simulated under JC69 (black), HKY (red), and GTR+I+Γ (green) models, respectively.
Figure 27

(a) Rate of correct decision (100 tests)

(b) Rate of correct decision (100 tests)

(c) Rate of correct decision (100 tests)

(d) Rate of correct decision (100 tests)

(e) Rate of correct decision (100 tests)

(f) Rate of correct decision (100 tests)

(g) Rate of correct decision (100 tests)

(h) Rate of correct decision (100 tests)

(i) Rate of correct decision (100 tests)

(j) Rate of correct decision (100 tests)

(k) Rate of correct decision (100 tests)

(l) Rate of correct decision (100 tests)

(m) Rate of correct decision (100 tests)

(n) Rate of correct decision (100 tests)

(o) Rate of correct decision (100 tests)
with “longer” branch lengths is slightly higher, especially when the sample size is small (around 5,000 bp). Thus, including more coalescent independent sites improves the accuracy of the test. Based on our simulations, around 10,000 bp for both long and short branch lengths are sufficient to ensure 95% accuracy when the data consist of coalescent independent sites, and 90% accuracy when multi-loci DNA data is used. Notably, the ability to identify the root of symmetric species trees does not depend on the sample size (Figure 27 (m), (n), and (o)), since the accuracy of identifying the root of symmetric species trees only relates to the significance levels that we selected for the hypothesis tests. The effects of sample size are not surprising, since the site pattern probabilities are estimated more accurately with more coalescent independent sites or longer DNA sequences, which is helpful in estimating the evolutionary relationships.

4.3.2 Application to an eight-taxa North American rattlesnake data set

The simulation studies above show good accuracy and efficiency of the rooting method in identifying the root of a four-taxon species quartet. The next step is to examine the performance of our method for a larger empirical data set. We choose as a test case a data set of North American rattlesnakes that consists of samples from three subspecies of *Sistrurus catenatus* (*S. c. catenatus*, *S. c. edwardsii*, and *S. c. tergeminus*), three subspecies of *Sistrurus miliarius* (*S. m. miliarius*, *S. m. barbouri*, and *S. m. streckeri*), and two outgroups (*Agkistrodon contortrix* and *Agkistrodon piscivorus*). This is a multi-locus DNA data set with 19 genes and a total of 8466 base pairs. One individual is selected from each taxon to estimate the species tree and the root position. The estimated species tree is shown in Figure 28a, which is consistent with earlier analyses of Kubatko et al. (2011) and Chifman and Kubatko
(2014). With two known outgroups, *A. contortrix* and *A. piscivorus*, the putative root position is labeled in red lines in Figure 28a.

When the outgroups are unknown, the unrooted 8-taxon species tree estimated by SVQQuartets is shown in Figure 28b, with each branch labeled from 1 to 13. A total of 70 species quartets within this species tree are examined to explore the root position based on our method, and the scores described in the “Methods” section are recorded for each branch (Table 3, “8-taxon”). We also removed the two outgroup species and tested our method with the remaining six taxa (Figure 28c), and record the scores in Table 3 (6-taxon). Note that the branches of the six-taxon species tree are given the same label as in the eight-taxon species tree (Figure 28b). Thus, branches 1, 2, and 9 no longer exist in Figure 28c.

From the scores of each branch (Table 3), it is easy to see that branch 9 should be selected as the root position for the eight-taxa species tree, which is consistent with previous analyses (Figure 28a). When the outgroups are removed from the analysis, our method can still accurately determine the root position on branch 10 and 12 (Table 3, “6-taxon”). Note that every single test of the 70 species quartets in the eight-taxon species tree correctly determined the root position, indicating an extremely high power for our method.

4.4 Discussion

In this study, we develop a new method for rooting species-level phylogenies using site pattern probabilities. More specifically, our method roots the quartet species trees under the coalescent model, and then applies the results of rooted quartets to infer the root location in larger species trees. The accuracy of this method is
Figure 28: Application of the rooting method to the rattlesnake data set. (a) The 8-taxon species tree rooted by outgroups: *A. contortrix* and *A. piscivorus*. The inferred root position is labeled by a red line. (b) The unrooted 8-taxon species tree, with each branch labeled from 1 to 13. The root position indicated by our method is labeled in red. (c) The unrooted 6-taxon species tree (with outgroups removed), with each branch labeled as in (b). The rooting position indicated by our method is labeled in red.
Table 3: Rooting results for the North American rattlesnake data set

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-taxon</td>
<td>9.67</td>
<td>9.67</td>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.33</td>
<td>31.00</td>
<td>7.42</td>
<td>1.75</td>
<td>7.42</td>
<td>1.75</td>
</tr>
<tr>
<td>6-taxon</td>
<td>NA</td>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.33</td>
<td>NA</td>
<td>10.84*</td>
<td>1.42</td>
<td>NA*</td>
<td>1.42</td>
</tr>
</tbody>
</table>

*For the 6-taxon species tree, the root position lies on the branch “10+12” (Figure 28). The score for this branch is recorded under branch 10 in this table, and the score for branch 12 is thus labeled “NA”.

examined by simulation studies and by application to an empirical North American rattlesnake data set. Notably, our method for root phylogenetic trees does not require specification of an outgroup, which makes it acceptable under very general conditions.

4.4.1 Rooting phylogenetic trees under different nucleotide substitution models

For a given species tree, the probability distribution of all possible site patterns can be computed for different nucleotide substitution models (e.g., JC69, HKY85, GTR+I+Γ, etc.). Specifically, for the simplest model, JC69, the identical base frequencies and the constant nucleotide substitution rate produce identical site pattern probabilities in many cases. For instance, given a four-taxon tree, there are only 15 unique site pattern probabilities under the JC69 model (Chifman and Kubatko, 2014, 2015). That is to say, the site patterns that fall into the same category have identical probability, thus it is straightforward to use the mean of the site pattern probabilities within the same category to compute the test statistics we propose here.

More complex nucleotide substitution models, such as the HKY85 and the GTR+I+Γ models, etc., can be specified by setting different rates for nucleotide changes. For example, HKY85 allows base frequencies to be unequal and considers one transition
(substitutions between the two purines, A and G, or between the two pyrimidines, C and T) rate and one transversion (substitutions between a purine and a pyrimidine) rate, while the GTR model also allows unequal base frequencies, but defines a symmetric parameter-rich substitution matrix. Under these complex nucleotide substitution models, there will be a larger number of distinct site pattern probabilities and computing the probability of any site pattern probability will be more complex compared to the JC69 model. Indeed, the site pattern probability under the coalescent cannot be expressed as an analytic expression for the GTR+I+Γ models, for example. However, the SVDQuartets method that is based on site pattern probabilities can still be applied to estimate a phylogenetic tree under models like HKY85 and GTR+I+Γ (Chifman and Kubatko, 2014, 2015), and it is not difficult to show that our rooting method can be applied to phylogenetic data under these complex nucleotide substitution models, as well. Although there are no explicit formulas and the site pattern probabilities may not be identical within the 15 categories described here, the relationship between site pattern categories $yxxx$ and $xyxx$, and between categories $xxyx$ and $xxxy$, for example, will still hold. We have simulated sequence data under both the HKY85 and GTR+I+Γ models in our simulation studies to verify that our method still applies under these complex nucleotide substitution models. Our results (Figure 27) indicate that the method works equally well under the three different nucleotide substitution models, regardless of the equality of base frequencies and substitution rates between bases.
4.4.2 Rooting phylogenetic trees using multi-loci DNA sequence data

Note that our rooting method assumes free recombination among the sites. In other words, it is designed for coalescent independent sites. However, previous simulation studies and real-data analyses also indicated good performance of SVDQuartets in analyzing multi-loci DNA sequence data. Also, SVDQuartets is suitable for the case of variable substitution rates across sites (i.e., substitution rates drawn from an arbitrary Gamma distribution) (Yang, 1993, 1994). The conclusion is similar for the rooting method presented here. As shown in Figure 27, the method is highly accurate in identifying the root positions when varying substitution rates are drawn from an arbitrary Gamma distribution. Furthermore, the simulation studies that simulate multi-loci DNA sequence data also show good performance. This is quite reasonable, because under the coalescent model, the distribution of expected gene trees across loci for multi-loci DNA sequence data should be consistent with that obtained for independent sites, and thus the site pattern frequency distribution should be close to one another when each gene has a similar size. From Figure 27, when there are more than 100 genes (10,000 bp in total), multi-loci DNA sequence data can be safely used to estimate rooted species tree directly from the site pattern probabilities.

4.4.3 Control of familywise error rate

Controlling the familywise error rate appropriately when performing multiple hypotheses tests is a well-studied topic. In our method, we considered two hypothesis tests at the same time. To ensure a 95% confidence level, we choose to control the total Type I error at level 0.05. Using the Bonferroni correction (Bonferroni, 1936),
the significance level for each test is selected to be 0.025 in all of our simulation studies. Based on the hypothesis tests, when neither test can be rejected, we infer the symmetric species tree. Thus, the probability that a symmetric tree is inferred when the tree is indeed symmetric should exceed 95%, since the Bonferroni test is conservative when the test are not independent, as is the case here. Figure 27 (m) - (o) shows the results of correctly identifying the symmetric species tree. Obviously, with coalescent independent sites, the power of the tests is right about 95% on average, while for multi-loci DNA sequence data, the power of the tests is slightly lower than 95%, with larger variance. This can be explained by the violation of free recombination for multi-loci DNA sequence data. When each nucleotide is not independent from each other, it is reasonable for us to observe a larger variance and a slightly lower power. In general, even with multi-loci DNA sequence data, the power of our rooting method still exceeds 90%, indicating that this rooting method is an accurate and efficient way to locate the root position in a species tree.

Setting the significance level at 0.025 for both tests gives very good performance in all of our simulation studies. However, choosing different significance levels is also possible. In fact, we recommend that users select larger significant levels with a small sample size, and choose smaller significant levels with a huge data set. The relationship between margin of error and sample size is well-studied (Altman and Bland, 2005; Rusticus and Lovato, 2014). Generally, larger sample sizes will lead to lower p-values (Sullivan and Feinn, 2012; Lin et al., 2013), thus requiring a smaller significance level. Additionally, the significance levels of the two hypothesis tests are not required to be identical. Once the sum of the two tests is smaller than 0.05, the overall error rate will be controlled at 5%. Thus, in general, differing significance levels
can be picked for each test, depending on the relative importance for the application of interest.
Chapter 5: Discussion and Future Work

Currently, coalescent theory is widely used in phylogenetic tree inference, and a high level of incongruence between gene trees and the species tree, as well as among estimated gene trees, is widely observed in empirical studies (Degnan and Salter, 2005; Degnan and Rosenberg, 2006; Pollard et al., 2006; Liu and Pearl, 2007; Degnan and Rosenberg, 2009; Spinks and Shaffer, 2009; Betancur-R et al., 2013; Salichos and Rokas, 2013; Pyron et al., 2014; Edwards et al., 2016). Though some other causes such as rate variation among genes, gene flow/hybridization, natural selection or potential estimation error may cause this incongruence (Avise et al., 1987; Rand, 2001; Sanderson and Shaffer, 2002; Funk and Omland, 2003; Ballard and Whitlock, 2004; Ballard and Rand, 2005; Spinks and Shaffer, 2009; Simmons et al., 2016), our research suggests that ILS plays an important role in causing gene tree - species tree and gene tree - gene tree conflict. To better understand the implications of ILS and to utilize this understanding to improve phylogenetic tree inference methods, we explore and highlight the extent of gene tree incongruence caused by ILS, quantitatively measure the gene tree incongruence based on the RF distances, establish a coalescent model that incorporates gene flow, and develop an efficient species tree rooting method by establishing a series of hypothesis tests to estimate rooted species trees. Overall, our research contributes to both statistical modeling and empirical phylogenetic research,
and it may benefit numerous evolutionary studies that require a good understanding of gene tree incongruence, the effects of gene flow, or identifying the evolutionary ancestor in the absence of an outgroup.

In Chapter 2, we discuss the importance of considering gene tree conflict in studies that use phylogenies, and regarding gene tree incongruence caused by ILS as an intrinsic part of a phylogeny (Maddison, 1997). Actually, failure to consider gene tree incongruence generated by ILS may cause severe biases in the phylogenetic inference results. In our study, the extent of expected gene tree incongruence introduced by ILS is quantified in two ways: the probability of observing two distinct sampled gene trees under the same species tree, and the probability distribution of RF distances for specific species trees. Though our study is focused on cases of small species trees (with fewer than 8 taxa), the trend obviously shows that the probability that two sampled gene trees match can be appreciably low. Thus, any phylogenetic inference method based on gene trees should not simply assume that the phylogenies are topologically congruent. To make a more confident conclusion, a thorough study that accounts for variation in the possible gene tree topologies is necessary. To help people better understand the potential gene tree incongruence cause by ILS, we provide a set of scripts that can be used with COAL (Degnan and Salter, 2005) and R to compute gene tree probabilities and RF distance distributions, as well as excel files that will compute the probability of sampling two matching gene trees from asymmetric species trees with three to eight taxa for a set of user-specified branch lengths (https://github.com/tian52/2016Match/tree/master/Scripts). Before concluding that gene tree incongruence is due to other causes, we suggest that researchers use these tools first to determine the expected gene tree incongruence, and
thus to assess whether ILS is a reasonable explanation for the extent of observed gene tree incongruence.

The ability to compute the gene tree probability distribution under the coalescent model greatly enhances our understanding of species tree inference (Degnan and Salter, 2005). However, previous studies on gene tree probability distribution only consider the coalescent in species tree - gene tree incongruence. In fact, gene flow may become an important factor that influences the probability distribution of the gene trees. In Chapter 3, we present a method for computing the gene tree probability distribution under the coalescent model for three species that allows gene flow between both pairs of sister populations. More significantly, this method is not limited in computing the gene tree topology distribution, the gene tree history distribution can also be computed. Our method leads to a better understanding of evolutionary relationships among closely related populations or species, with high potential for gene flow. Applying this method, the computed coalescent history distributions can be used to infer species tree parameters, such as the ancestral effective population sizes and the gene flow rates. In Chapter 3, we use both simulation studies and empirical data sets to infer these parameters under a maximum likelihood framework. A very interesting finding of this study is that, for a fixed species tree topology, many different coalescent parameter values may lead to the same distribution on gene tree topologies, while the distribution of the coalescent histories is distinct for different choices of parameters, thus expanding our knowledge of the identifiability of coalescent parameters along a fixed species tree. In the presence of gene flow, only the distribution of gene tree topologies alone are not enough to infer species tree coalescent parameters.
In Chapter 4, we developed a computationally efficient phylogenetic tree rooting method to estimate the rooted phylogenetic trees using a series of hypothesis tests. The hypothesis tests are established from algebraic statistical methods. More specifically, the probability distribution of all possible site patterns under a given species tree is computed to derive the hypothesis tests. The process of deriving the hypothesis tests from the simplest JC69 nucleotide substitution model is presented in Chapter 4. Furthermore, simulation studies show that our method also work for more complex nucleotide substitution models, such as the HKY85, and GTR+I+Γ model, etc. One limitation of using the rooting method to root phylogenetic trees is that our rooting method assumes free recombination among the sites. However, from simulation studies and real-data analyses, it is still safe to use this rooting method in analyzing multi-loci DNA sequence data. Overall, the rooting method is highly accurate in identifying the root positions in both simulation studies and empirical data sets. Note that the sample size (number of coalescent independent sites, number of genes and sequence lengths) should not be too small to ensure that the method has high power. In sum, the rooting method provides a computationally practical method to efficiently root species trees under the coalescent model. Importantly, the new rooting method will benefit numerous evolutionary studies which require rooting a phylogenetic tree without having to specify any outgroups (Reich et al., 2011; Aguinaldo et al., 1997; Philippe et al., 2011; Waddell et al., 1999; Madsen et al., 2001; Murphy et al., 2001; Scally et al., 2001; Cox et al., 2008; Lake, 2009).

Development of phylogenetic tree inference methods under the coalescent model is an active area of research, and there are many potential study directions based on our research. First of all, we suggest that researchers should use the expected
gene tree incongruence introduced in Chapter 2 to assess whether ILS is a reasonable explanation for the extent of observed gene tree incongruence before explaining the gene tree incongruence by other causes. To correctly deal with the possibility of gene tree incongruence, incorporating phylogenetic uncertainty in the underlying trees for all traits under consideration by integrating over all possible gene trees under a particular evolutionary model is a good solution, though with a substantial computational cost. Second, an interesting project is to extend the coalescent model with gene flow (described in Chapter 3) to an arbitrary number of species with more than one sampled gene per species. Such studies will greatly extend our scope in understanding the effects of gene flow under the coalescent. Last but not the least, because of the efficiency and the accuracy of our rooting method in Chapter 4, it is worth implementing this rooting method into some software packages like PAUP* (Swofford, 2001), which already includes methods to estimate unrooted phylogenetic species trees. There will be a more extensive use of the computational feasible rooting method once it is implemented into PAUP*. There is no doubt that these potential future research directions will expand our understanding in phylogenetic species tree inference under the coalescent model.
Bibliography


Appendix A: Derivation of the gene tree history distribution 
under the coalescent model

For G1H4, the first coalescent event occurs after time $\tau_1$, while the second coalescent event occurs after time $\tau_2$, and thus $\tau_1 < t_1 < \tau_2 < t_2$. From the present to time $\tau_1$, no coalescent event occurs, and the probability is

$$ P(t_1 > \tau_1) = 1 - \int_0^{\tau_1} \left[ c_1(e^{Q_1t_1})_{21} + c_2(e^{Q_1t_1})_{23} \right] dt_1. \quad (A.1) $$

Similar to history G1H3, after time $\tau_1$, there are four distinct ways in which the two coalescent events can happen. In all four cases, the process starts in state ddc. The first case, denoted by G1H4C1, goes from state ddc to ddd, and then a coalescent event occurs and the state is dd. The final coalescent event does not occur before time $\tau_2$. Thus the state can change to any of the three states: dd, dc, and cc. To model this, we use $Q_3$ to calculate the change from state ddc to ddd, and use $Q_2$ for the change from state dd to dd, dc or cc. The density function is

$$ f_{G1H4C1}(t_2) = c_4(e^{Q_3(t_2-\tau_1)})_{21}[(e^{Q_2(\tau_2-t_2)})_{11} + (e^{Q_2(\tau_2-t_2)})_{12} + (e^{Q_2(\tau_2-t_2)})_{13}] . \quad (A.2) $$

Similarly, we can write the density functions for the other three probabilities. The second probability has the sequence of states ddc - ddc - dc - dd/dc/cc, with corresponding density function

$$ f_{G1H4C2}(t_2) = c_4(e^{Q_3(t_2-\tau_1)})_{22}[(e^{Q_2(\tau_2-t_2)})_{21} + (e^{Q_2(\tau_2-t_2)})_{22} + (e^{Q_2(\tau_2-t_2)})_{23}] . \quad (A.3) $$
The third probability has the sequence of states ddc - ccc - cc - dd/dc/cc, with corresponding density function

\[ f_{G1H4C3}(t_2) = c_3(e^{Q3(t_2-\tau_1)})_{28}[(e^{Q2(\tau_2-t_2)})_{31} + (e^{Q2(\tau_2-t_2)})_{32} + (e^{Q2(\tau_2-t_2)})_{33}]. \quad (A.4) \]

Finally, the fourth probability has the sequence of states ddc - ddc - dc - dd/dc/cc, with corresponding density function

\[ f_{G1H4C4}(t_2) = c_3(e^{Q3(t_2-\tau_1)})_{27}[(e^{Q2(\tau_2-t_2)})_{21} + (e^{Q2(\tau_2-t_2)})_{22} + (e^{Q2(\tau_2-t_2)})_{23}]. \quad (A.5) \]

Thus the overall density function for \( t_1 \) and \( t_2 \) for history G1H4 is

\[ f_{G1H4}(t_2) = f_{G1H4C1}(t_2) + f_{G1H4C2}(t_2) + f_{G1H4C3}(t_2) + f_{G1H4C4}(t_2), \quad (A.6) \]

and the marginal probability for G1H4 is

\[ P(G1H4) = P(t_1 > \tau_1) \left( \int_{\tau_1}^{t_2} f(G1H4, t_2) dt_2 \right) \quad (A.7) \]

For G1H5, both coalescent events occur after time \( \tau_2 \) and the lineages that come from species A and B coalesce first, thus \( \tau_2 < t_1 < t_2 \). After time \( \tau_2 \), any two lineages have the same probability of coalescing. Since no coalescent events occur before time \( \tau_1 \), after time \( \tau_1 \) the state can change from ddc to all eight possible states. Thus, the probability of G1H5 is

\[ P(G1H5) = \frac{1}{3} P(t_1 > \tau_1)[(e^{Q3(\tau_2-\tau_1)})_{21} + (e^{Q3(\tau_2-\tau_1)})_{22} + (e^{Q3(\tau_2-\tau_1)})_{23} + (e^{Q3(\tau_2-\tau_1)})_{24} + (e^{Q3(\tau_2-\tau_1)})_{25} + (e^{Q3(\tau_2-\tau_1)})_{26} + (e^{Q3(\tau_2-\tau_1)})_{27} + (e^{Q3(\tau_2-\tau_1)})_{28}] \quad (A.8) \]

History G2H1 can be analyzed with a procedure similar to that used for history G1H3, and the probability of G2H1 is

\[ P(G2H1) = P(t_1 > \tau_1) \left( \int_{\tau_1}^{t_2} \int_{\tau_1}^{t_2-t_1} f(G2H1, t_1, t_2) dt_2 dt_1 \right). \quad (A.9) \]
Here

\[ f_{G2H1}(t_1, t_2) = f_{G2H1C1}(t_1, t_2) + f_{G2H1C2}(t_1, t_2) + f_{G2H1C3}(t_1, t_2) + f_{G2H1C4}(t_1, t_2). \]  

(A.10)

In the case of G2H1, some of the state changes are different than G1H3. The first case still goes from the state ddc to ddd, with corresponding density function

\[ f_{G2H1C1}(t_1, t_2) = c_4(e^{Q_3(t_1-\tau_1)})_{21}[c_4(e^{Q_2(t_2-\tau_1)})_{11} + c_3(e^{Q_2(t_2-\tau_1)})_{13}] . \]  

(A.11)

The second case has the sequence of states ddc - dcc - dc - dd/cc, with corresponding density function

\[ f_{G2H1C2}(t_1, t_2) = c_3(e^{Q_3(t_1-\tau_1)})_{25}[c_4(e^{Q_2(t_2-\tau_1)})_{21} + c_3(e^{Q_2(t_2-\tau_1)})_{23}] . \]  

(A.12)

The third probability has the sequence of states ddc - ccc - cc - dd/cc, with corresponding density function

\[ f_{G2H1C3}(t_1, t_2) = c_3(e^{Q_3(t_1-\tau_1)})_{28}[c_4(e^{Q_2(t_2-\tau_1)})_{31} + c_3(e^{Q_2(t_2-\tau_1)})_{33}] . \]  

(A.13)

The last probability has the sequence of states ddc - dcc - dc - dd/cc, with corresponding density function

\[ f_{G2H1C4}(t_1, t_2) = c_4(e^{Q_3(t_1-\tau_1)})_{24}[c_4(e^{Q_2(t_2-\tau_1)})_{21} + c_3(e^{Q_2(t_2-\tau_1)})_{23}] . \]  

(A.14)

Similar to history H1G4, we can write the probability of history G2H2:

\[ f_{G2H2C1}(t_2) = c_4(e^{Q_3(t_2-\tau_2)})_{21}[(e^{Q_2(\tau_2-t_2)})_{11} + (e^{Q_2(\tau_2-t_2)})_{12} + (e^{Q_2(\tau_2-\tau_1-t_2)})_{13}] . \]  

(A.15)

\[ f_{G2H2C2}(t_2) = c_3(e^{Q_3(t_2-\tau_1)})_{25}[(e^{Q_2(\tau_2-t_2)})_{21} + (e^{Q_2(\tau_2-t_2)})_{22} + (e^{Q_2(\tau_2-t_2)})_{23}] . \]  

(A.16)
\[ f_{G2H2C3}(t_2) = c_3(e^{Q_3(t_2-\tau_1)})_{28}[(e^{Q_2(t_2-t_1)})_{31} + (e^{Q_2(t_2-t_1)})_{32} + (e^{Q_2(t_2-t_1)})_{33}]. \quad (A.17) \]

\[ f_{G2H2C4}(t_2) = c_4(e^{Q_3(t_2-\tau_1)})_{24}[(e^{Q_2(t_2-t_1)})_{21} + (e^{Q_2(t_2-t_1)})_{22} + (e^{Q_2(t_2-t_1)})_{23}]. \quad (A.18) \]

\[ f_{G2H2}(t_2) = f_{G2H2C1}(t_2) + f_{G2H2C2}(t_2) + f_{G2H2C3}(t_2) + f_{G2H2C4}(t_2). \quad (A.19) \]

\[ P(G2H2) = P(t_1 > \tau_1) \left( \int_{\tau_1}^{\tau_2} f(G2H2, t_2) \, dt_2 \right) \quad (A.20) \]

When both coalescent events occur after time \( \tau_2 \), any two lineages will have the same probability of coalescing, thus the probability of history G2H3 should be exactly the same as history G1H5. Also, due to the symmetry between G3H1 to G3H3 and G2H1 to G2H3, the probabilities of G3H1, G3H2, and G3H3 should be equal to the probabilities of G2H1, G2H2, and G2H3, respectively. Thus we have

\[ P(G2H3) = P(G1H5) \quad (A.21) \]

\[ P(G3H1) = P(G2H1) \quad (A.22) \]

\[ P(G3H2) = P(G2H2) \quad (A.23) \]

\[ P(G3H3) = P(G2H3) \quad (A.24) \]