2
3 Distribution of coalescent histories under the coalescent model with
4 gene flow
56 Yuan Tian¹, Laura S. Kubatko^{1,2*}
7

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

23 Key words: gene tree, species tree, coalescent model, gene flow, migration, topology, coalescent
24 history

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 (see 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 length of the speciation intervals (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 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 histories. 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 the 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.

Figure 1: The model species tree $((A, B), C)$ for three species with gene flow between sister species. The speciation times are denoted as τ_1 and τ_2 , respectively. θ_A , θ_B , θ_C , θ_{AB} , and θ_{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 .

2 MATERIAL AND METHODS

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 1). The two ancestral species are denoted AB and ABC. The time since speciation occurred between A and B is denoted τ_1 , and the time since the speciation event between AB and C is denoted τ_2 , where τ_1 and τ_2 are measured by the expected number of mutations per site.

138 The genetic data that we will analyze contain multiple loci. For every locus, we assume that one
 139 sequence was sampled from each of the three species. It is assumed that there is no recombination
 140 within a locus, and free recombination among loci. There are three possibilities for the gene tree
 141 topology relating the three sampled sequences at a locus: $((A, B), C)$, $((B, C), A)$, and $((A, C),$
 142 $B)$. Probability distributions relating to the coalescent history can be derived using Markov chains
 143 based on the structured coalescent process, as in Hobolth et al. (2011).

Figure 2: (a) Five possible coalescent histories with gene tree topology $((A, B), C)$ (denoted by G1). For these coalescent histories, the first (most recent) coalescent event always occurs between the lineages from species A and B. If the first coalescent event occurs before τ_1 , and the second coalescent event occurs between τ_1 and τ_2 , the gene tree has coalescent history G1H1. If the first coalescent event occurs before τ_1 , but the second coalescent event occurs after τ_2 , the gene tree has coalescent history G1H2. The other three coalescent histories are denoted by G1H3, G1H4, and G1H5. (b) Three possible coalescent histories with gene tree topology $((B, C), A)$ (denoted by G2). (c) Three possible coalescent histories with gene tree topology $((A, C), B)$ (denoted by G3). For the coalescent histories in (b) and (c), the first coalescent event occurs between the lineages from species B and C (for topology G2) or species A and C (for topology G3).

144 In our model, we consider gene flow between sister species A and B, and between sister species
 145 AB and C. More specifically, for species A and B, gene flow between contemporaneous populations
 146 can occur during the time period from the present to time τ_1 ; for species AB and C, gene flow
 147 can occur between times τ_1 and τ_2 (assume that there is no gene flow between species B and C, or
 148 between species A and C, after time τ_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 τ_2 (see Figure 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 μ 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 2).

It is easy to see that high rates of gene flow will generate more variation in coalescent histories. Under our model in Figure 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 2(a), G1H1 to G1H5 show five different coalescent histories consistent with topology $((A, B), C)$. In Figure 2(b), G2H1 to G2H3 show three coalescent histories consistent with topology $((B, C), A)$. G3H1 to G3H3 are shown in Figure 2(c). The speciation times are labeled as τ_1 and τ_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 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 θ_x, c_x, m_x , and τ_x as the original parameters, which are not scaled by $2/\theta_0$. For the scaled parameters, we

181 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
 182 parameters, and speciation times, respectively. More details can be found in the Theory section.

183 2.2 Maximum likelihood parameter estimation

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

190 Three simulation studies were carried out using the software ms (Hudson, 2002) and seq-gen
 191 (Rambaut and Grassly, 1997). The ms software was used to simulate gene trees directly (for the
 192 first simulation study), while seq-gen was used to simulate 500bp and 1000bp DNA sequences
 193 on the simulated gene trees under the JC69 model (Jukes and Cantor, 1969) for the second and
 194 third simulation studies, respectively. The DNA sequence data sets were analyzed with PAUP*
 195 (Swofford, 2003) to estimate gene trees under the maximum likelihood criterion. In each simulation
 196 study, we selected a varying number of loci (ranging from 50 to 100,000) to assess our model. All
 197 data were simulated under the fixed species tree ((A, B), C), with $\theta_A = \theta_B = \theta_C = \theta_{AB} = 0.005$,
 198 $m_1 = m_2 = 200$, $\tau_1 = 0.004$, and $\tau_2 = 0.006$, which were chosen based on Zhu and Yang (2012).
 199 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
 200 $T_2 = 2.4$.

201 For each simulated data set with K loci, the frequency of the x^{th} coalescent history was recorded
 202 as k_x , $x = 1, 2, \dots, 11$. In order to estimate the model parameters, we consider two methods for
 203 searching parameter space to find the maximum likelihood estimate (MLE), both based on a grid
 204 search. The first assumes that $M_1 = M_2 = M$ and $C_1 = C_2 = C_3 = C_4 = C$, with C varying from
 205 0 to 2 (we used 200 equal spaced values), and M varying from 0.001 to 10 (we used 200 different
 206 values on the log scale). The other method assumes that $C_1 = C_2 = C_3 = C_4 = 1$, and varies both
 207 M_1 and M_2 from 0.001 to 10 (we used 200 different values on the log scale). Note that although we
 208 could consider M_1 , M_2 , and C at the same time, in the simulation study, we only considered two
 209 parameters at a time in order to reduce the computational burden and to run more replications.
 210 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 THEORY

3.1 The probability distribution of coalescent histories

In our proposed model (Figure 1), we divide the species tree into three time periods. The first goes from the present time to τ_1 , and three species (A, B, and C) exist during this time period; the second time period goes from τ_1 to τ_2 , with two species AB and C; and the last goes from time τ_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.

3.1.1 The instantaneous rate matrix for each time period.

During the first time period, from the present to τ_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 5×5 instantaneous rate matrix Q1:

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

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 $e^Q = \sum_{i=0}^{\infty} Q^i/i!$. The $(j, k)^{th}$ entry of e^Q is denoted as $(e^Q)_{jk}$. Note that $(e^{Q^t})_{jk}$ represents **probability that the system starts in state j and is in state k at time t** .

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 τ_1 is

$$f(t_1) = c_1(e^{Q1t_1})_{21} + c_2(e^{Q1t_1})_{23}, \quad t_1 < \tau_1. \quad (1)$$

When the first coalescent event occurs before time τ_1 ($t_1 < \tau_1$), a matrix Q2 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 τ_1 now become dd, dc, cc, d, c, and the associated rate matrix is

$$Q2 = \begin{matrix} & \begin{matrix} dd & dc & cc & d & c \end{matrix} \\ \begin{matrix} dd \\ dc \\ cc \\ d \\ c \end{matrix} & \begin{pmatrix} -2m_2 - c_4 & 2m_2 & 0 & c_4 & 0 \\ m_2 & -2m_2 & m_2 & 0 & 0 \\ 0 & 2m_2 & -2m_2 - c_3 & 0 & c_3 \\ 0 & 0 & 0 & -m_2 & m_2 \\ 0 & 0 & 0 & m_2 & -m_2 \end{pmatrix} \end{matrix}$$

During the time period from τ_1 to τ_2 , the probability density for the coalescent event at time $t_2 < \tau_2$ is

$$f(t_2) = c_4(e^{Q2(t_2-\tau_1)})_{21} + c_3(e^{Q2(t_2-\tau_1)})_{23}, \quad t_2 < \tau_2. \quad (2)$$

The system becomes more complicated if the first coalescent event occurs after time τ_1 ($t_1 > \tau_1$). In that case, three lineages exist after time τ_1 and all of them can migrate between species AB and C. A 13×13 rate matrix Q3 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,

263

264

265

$$Q3 = \begin{matrix} & \begin{matrix} ddd & ddc & dcd & cdd & dcc & cdc & ccd & ccc & dd & dc & cc & d & c \end{matrix} \\ \begin{matrix} ddd \\ ddc \\ dcd \\ cdd \\ dcc \\ cdc \\ ccd \\ ccc \\ dd \\ dc \\ cc \\ d \\ c \end{matrix} & \left(\begin{array}{cccccccccccc} -- & m_2 & m_2 & m_2 & 0 & 0 & 0 & 0 & 3c_4 & 0 & 0 & 0 & 0 \\ m_2 & -- & 0 & 0 & m_2 & m_2 & 0 & 0 & 0 & c_4 & 0 & 0 & 0 \\ m_2 & 0 & -- & 0 & m_2 & 0 & m_2 & 0 & 0 & c_4 & 0 & 0 & 0 \\ m_2 & 0 & 0 & -- & 0 & m_2 & m_2 & 0 & 0 & c_4 & 0 & 0 & 0 \\ 0 & m_2 & m_2 & 0 & -- & 0 & 0 & m_2 & 0 & c_3 & 0 & 0 & 0 \\ 0 & m_2 & 0 & m_2 & 0 & -- & 0 & m_2 & 0 & c_3 & 0 & 0 & 0 \\ 0 & 0 & m_2 & m_2 & 0 & 0 & -- & m_2 & 0 & c_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & m_2 & m_2 & m_2 & -- & 0 & 0 & 3c_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -- & 2m_2 & 0 & c_4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & m_2 & -- & m_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2m_2 & -- & 0 & c_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -- & m_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & m_2 & -- \end{array} \right) \end{matrix}$$

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

In our model, we use Q3 for the time period between time τ_1 and t_1 when the first coalescent event occurs after time τ_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.

3.1.2 The probability distribution of the coalescent 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), and ((A, C), B), respectively (Figure 2). 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 2a). 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

281 Appendix.

282 For G1H1, the first coalescent event occurs between the present and time τ_1 , while the second
 283 coalescent event occurs between time τ_1 and τ_2 , thus $t_1 < \tau_1 < t_2 < \tau_2$. To derive the joint density
 284 of the coalescent times for this coalescent history, we first consider the time range between the
 285 present and time τ_1 . Since there is no gene flow between species C and any other species during
 286 this time, we only need to calculate the probability that the two lineages in species A and B coalesce
 287 before time τ_1 . Due to the possibility of gene flow between species A and B, the coalescent event
 288 can happen in either species A or B. The whole process can be described as three lineages that
 289 start in state abc, and then right before the first coalescent event, the state is either aac or bbc.
 290 Thus from (1), we can derive

$$f_{G1H1}(t_1) = c_1(e^{Q1t_1})_{21} + c_2(e^{Q1t_1})_{23}, \quad t_1 < \tau_1 \quad (3)$$

291 Similarly, the second coalescent event occurs between time τ_1 and τ_2 , and the process is assumed
 292 to start from state dc. Right before the second coalescent event, the state is dd or cc. From the
 293 previous function (2), we have

$$f_{G1H1}(t_2) = c_4(e^{Q2(t_2-\tau_1)})_{21} + c_3(e^{Q2(t_2-\tau_1)})_{23}, \quad \tau_1 < t_2 < \tau_2 \quad (4)$$

294 The joint distribution for both t_1 and t_2 is then

$$\begin{aligned} f_{G1H1}(t_1, t_2) &= f_{G1H1}(t_1) \times f_{G1H1}(t_2) \\ &= [c_1(e^{Q1t_1})_{21} + c_2(e^{Q1t_1})_{23}][c_4(e^{Q2(t_2-\tau_1)})_{21} + c_3(e^{Q2(t_2-\tau_1)})_{23}], 0 < t_1 < \tau_1 < t_2 < \tau_2. \end{aligned} \quad (5)$$

295 To find the marginal probability of coalescent history G1H1, we integrate out the gene tree coa-
 296 lescent times,

$$P(G1H1) = \int_0^{\tau_1} [c_1(e^{Q1t_1})_{21} + c_2(e^{Q1t_1})_{23}] dt_1 \int_{\tau_1}^{\tau_2} [c_4(e^{Q2(t_2-\tau_1)})_{21} + c_3(e^{Q2(t_2-\tau_1)})_{23}] dt_2. \quad (6)$$

297 For G1H2, the first coalescent event occurs between the present and time τ_1 , while the second
 298 coalescent event occurs after time τ_2 , and thus $t_1 < \tau_1 < \tau_2 < t_2$. The only difference in the
 299 calculations for G1H2 is that no coalescent event occurs between time τ_1 and τ_2 . The probability
 300 that the second coalescent event occurs before time τ_2 is

$$P(t_2 < \tau_2 | t_1 < \tau_1 < t_2) = \int_{\tau_1}^{\tau_2} [c_4(e^{Q2(t_2-\tau_1)})_{21} + c_3(e^{Q2(t_2-\tau_1)})_{23}] dt_2. \quad (7)$$

301 Thus, the probability that the second coalescent event occurs after time τ_2 is

$$P(t_2 > \tau_2 | t_1 < \tau_1 < t_2) = 1 - \int_{\tau_1}^{\tau_2} [c_4(e^{Q2(t_2-\tau_1)})_{21} + c_3(e^{Q2t_2})_{23}] dt_2, \quad (8)$$

302 and

$$\begin{aligned}
P(G1H2) &= P(t_1 < \tau_1)P(t_2 > \tau_2 | t_1 < \tau_1 < t_2) \\
&= \int_0^{\tau_1} [c_1(e^{Q1t_1})_{21} + c_2(e^{Q1t_1})_{23}] dt_1 \left(1 - \int_{\tau_1}^{\tau_2} [c_4(e^{Q2(t_2-\tau_1)})_{21} + c_3(e^{Q2(t_2-\tau_1)})_{23}] dt_2 \right).
\end{aligned}
\tag{9}$$

303 For G1H3, the first coalescent event occurs after time τ_1 , while the second coalescent event
304 occurs before time τ_2 , and thus $\tau_1 < t_1 < t_2 < \tau_2$. From the present to time τ_1 , no coalescent event
305 occurs, and the probability is

$$P(t_1 > \tau_1) = 1 - \int_0^{\tau_1} [c_1(e^{Q1t_1})_{21} + c_2(e^{Q1t_1})_{23}] dt_1. \tag{10}$$

306 After time τ_1 , things are more complicated. There are four distinct ways in which the two
307 coalescent events can happen. In all four cases, the process starts in state ddc, which means that
308 two lineages are in population d and one is in population c. The first case, denoted by G1H3C1,
309 goes from state ddc to ddd, and then a coalescent event occurs and the state is dd. The final
310 coalescent event can occur either from this state, or by changing to state cc. To model this, we use
311 Q3 to calculate the change from state ddc to ddd, and use Q2 for the change from state dd to dd
312 or cc. The joint density function is

$$f_{G1H3C1}(t_1, t_2) = c_4(e^{Q3t_1})_{21} [c_4(e^{Q2(t_2-\tau_1)})_{11} + c_3(e^{Q2(t_2-\tau_1)})_{13}]. \tag{11}$$

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

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

$$f_{G1H3C2}(t_1, t_2) = c_4(e^{Q3t_1})_{22} [c_4(e^{Q2(t_2-\tau_1)})_{21} + c_3(e^{Q2(t_2-\tau_1)})_{23}]. \tag{12}$$

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

$$f_{G1H3C3}(t_1, t_2) = c_3(e^{Q3t_1})_{28} [c_4(e^{Q2(t_2-\tau_1)})_{31} + c_3(e^{Q2(t_2-\tau_1)})_{33}]. \tag{13}$$

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

$$f_{G1H3C4}(t_1, t_2) = c_3(e^{Q3t_1})_{27}[c_4(e^{Q2(t_2-\tau_1)})_{21} + c_3(e^{Q2(t_2-\tau_1)})_{23}]. \quad (14)$$

Thus, the overall density function for t_1 and t_2 for coalescent history G1H3 is

$$\begin{aligned} f_{G1H3}(t_1, t_2) = & f_{G1H3C1}(t_1, t_2) + f_{G1H3C2}(t_1, t_2) \\ & + f_{G1H3C3}(t_1, t_2) + f_{G1H3C4}(t_1, t_2), \end{aligned} \quad (15)$$

and the marginal probability of G1H3 is

$$P(G1H3) = P(t_1 > \tau_1) \left[\int_{\tau_1}^{\tau_2} \int_{\tau_1}^{\tau_2-t_1} f(G1H3, t_1, t_2) dt_2 dt_1 \right]. \quad (16)$$

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

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. 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 τ_1 and τ_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 coalescent history in (6). Using matrix $Q1$ as an example, note that for $c_0 = 2/\theta_0$ we can write

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

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 .

4 RESULTS

4.1 Gene flow between sister species produces different distributions of coalescent 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 3). 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 3(a)), and current and ancestral populations have unequal effective population sizes (Figure 3(b)).

When there is no gene flow between either pair of sister species (Figure 3(a) $M_1 = 0, M_2 = 0$), only three coalescent histories are possible (G1H4, G1H5, and G2H3/G3H3; see Figure 2), 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 2) 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 3(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 3(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 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 2 have positive probability (Figure 3(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 3(a)), the distribution changes when these parameters differ across the species tree. In Figure 3(b), we used a set of species trees with scaled coalescent parameters

Figure 3: The probability distribution of the coalescent 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 coalescent history is denoted by a different color as shown in the figure. Coalescent 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 ancestral species AB ($M_1 = 0.5, M_2 = 0$); 4, equal rates of gene flow in both pairs of sister populations ($M_1 = M_2 = 0.5$).

$C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2$ along with all of the other parameter combinations in Figure 3(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 3(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 3(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 2). 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 S1(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 3(a), G1H4 is one of the dominant coalescent histories, while in Figure 3(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 S1). Comparing Figure 3(a) and Figure S1(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 3(a)(b) and Figure S1(a)(b)).

We also considered a second level for the rate of gene flow. In Figure S1(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 3 and Figure S1(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.

4.2 Variation in the coalescent 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 4 and Figures S2 - S5, 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 4(a)-(i), Figures S2, S3), 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 4(j)-(l), Figures S4, S5). For the speciation times, a set of relatively shorter speciation times [$T_1 = 1.6; T_2 = 2.4$ (Figure 4(a)-(f),(j)-(l), Figures S2, S4)] and a set of longer speciation times [$T_1 = 2; T_2 = 4$ (Figure 4(g)-(i), Figures S3, S5)] are used. Figure 5 and Figures S6 - S9 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 $G1=((A,B),C)$, the second column shows the distribution of the three coalescent histories with the other two topologies $G2/G3$, and the last column shows the distribution of the three possible gene tree topologies ($G1$, $G2$, 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 4(a)-(c)) or between species AB and C (i.e., $M_2 = 0.001$; Figure 5(a)-(c)), different coalescent histories will have positive probability compared with the case in which gene flow is present. For example, comparing

Figure 4: 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 C and the ancestral species AB, M_2 (x-axis; shown on a log scale). Panels (a), (d), (g), and (j) show the probabilities of the five coalescent histories (G1H1 - G1H5) with topology G1. Panels (b), (e), (h), and (k) show the probabilities of the three coalescent 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). **PZ indicates population size.**

Figure 4(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 5(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 4(c),(f); Figure 5(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 4 and Figure 5). 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 4(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 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 4(f) and (i)). This effect is even larger in Figure 5(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 4(d)-(f) and (j)-(l); Figure 5(d)-(f) and (j)-(l); Figure S2 - S9).

Figure 5: 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.**

4.3 Different distributions of coalescent 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 unrooted 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 6 shows eight species trees ($S1 - S8$) that all have the same set of scaled speciation times, $T_1 = 1.6$ and $T_2 = 2.4$. Trees $S1 - S4$ have equal effective population sizes with scaled coalescent parameters $C_1 = C_2 = C_3 = C_4 = 1$, while species trees $S5 - S8$ 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, $S1$ shows the case of very low gene flow ($M_1 = M_2 = 0.001$), while $S5$ 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: $S2$ has small gene flow rates ($M_1 = 0.5$; $M_2 = 0.544$); $S8$ has medium gene flow rates ($M_1 = 2$; $M_2 = 5$); and $S4$ 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 6 divide the total probability into the three probabilities corresponding to the gene tree topologies (from left to right, G1, G2, and G3), which are *identical* for all eight cases. Under this identical topology distribution ($P(G1) = 0.7$, $P(G2) = P(G3) = 0.15$), the distributions of

Figure 6: Coalescent history distributions for eight species trees (rows labeled $S1 - S8$) 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 $P(G1) = 0.7$, and $P(G2) = P(G3) = 0.15$. 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 $S1 - S4$, and $C_1 = 1, C_2 = C_3 = 0.5, C_4 = 0.2$ for species trees $S5 - S8$. All species trees have different rates of gene flow: $S1$: $M_1 = M_2 = 0$; $S2$: $M_1 = 0.5$, $M_2 = 0.544$; $S3$: $M_1 = 2$, $M_2 = 2.23$; $S4$: $M_1 = 20$, $M_2 = 28$; $S5$: $M_1 = 0.796$, $M_2 = 0$; $S6$: $M_1 = 1.337$, $M_2 = 0.5$; $S7$: $M_1 = 1.894$, $M_2 = 2$; $S8$: $M_1 = 2$, $M_2 = 5$.

487 coalescent histories are very different for these eight species trees.

488 Notably, these eight species trees are not the only species trees that share this particular gene
489 tree topology distribution, and this gene tree topology distribution is not the only one which can
490 be generated by multiple species trees. However, despite the implied non-identifiability of gene flow
491 based only on the topology distribution, we note that the coalescent history distributions appear to
492 be distinct in these cases, though we have not shown this formally. Because of this, the information
493 provided by the distribution of coalescent histories can be used to estimate the parameters (effective
494 population sizes and gene flow rates) in the coalescent model with gene flow.

495 *Simulation studies for maximum likelihood parameter estimation*

496 Based on these results, we developed a method to estimate the rates of coalescence and of gene
497 flow along a fixed species phylogeny using maximum likelihood (see Materials and Methods for
498 details). To assess the performance of our method, we carried out three simulation studies. In
499 the first simulation study, we simulated gene trees directly with a varying number of loci ranging
500 from 50 to 100,000. Gene trees were simulated under the fixed species tree ((A, B), C), with
501 $\theta_A = \theta_B = \theta_C = \theta_{AB} = 0.005$, $m_1 = m_2 = 200$, $\tau_1 = 0.004$, and $\tau_2 = 0.006$, corresponding to
502 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
503 that $M_1 = M_2 = M$ and $C_1 = C_2 = C_3 = C_4 = C$, we evaluated the likelihood of 40,000 species
504 trees (each differing in the values of M and C but with the topology and branch lengths fixed) to
505 find the **maximum likelihood estimates (MLEs)** of M and C . Figure 7 shows the results for one
506 simulated data set in this simulation study for which 1000 gene trees were simulated with $C = 1$
507 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
508 Figure 7, are close to the true values.

509 We repeated this simulation process 1,000 times for varying numbers of loci, and obtained the
510 mean and standard deviation for the MLEs of C and M (Figure 8(a), top two plots). We see
511 that the estimates of C and M appear to be generally unbiased, with increasing variance as the
512 number of loci decreases. We also consider the case in which we are reasonably confident that
513 $C_1 = C_2 = C_3 = C_4 = 1$ and wish to estimate M_1 and M_2 separately (Figure 8(a), bottom two
514 plots). Again, our results suggest very good performance of our method in estimating the rates of
515 gene flow in a three-species model, with unbiased estimates and decreasing variance as the number
516 of loci increases.

517 The first simulation study indicates good performance of our method when gene trees are known

Figure 7: 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 numerical issues when performing matrix exponential.

(a)

(b)

Figure 8: 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.

without error. However, in the typical empirical setting, gene trees must first be estimated from observed sequence data (Figure 8(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.

4.4 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. 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 θ , M_1 , and M_2 in a two-step procedure. In the first step, we examined 60 equally-spaced values for θ 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 θ ranging from 0.00396 to 0.01188, and four values of M_1 (0.01, 0.1, 1, and 10) (Figure 9). 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.

Figure 9: 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 $l = -1,189.50$.

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 4(c) and Figure 5(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 4(c) to Figures 4(f), (i), and (l); compare Figure 5(l) to Figures 5(c), (f), and (i)). This finding is sensible, 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.

This 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 6). 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.

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 mosquitos), 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 8). 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 θ_x and the speciation time τ_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 τ/θ 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.

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 29×29 , 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 1, our model assumes no gene flow between species A and C, and species B and C from the present to time τ_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 τ_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.

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.

6 ACKNOWLEDGEMENT

We thank Drs. Matthew Hahn and James Pease for their willingness to share the mosquito data, and Dr. Pease for assistance in providing the data. We thank Dr Luay Nakhleh for comments on an earlier version, and for his willingness to share the sampled mosquito data. We thank anonymous reviewers who provided helpful comments on earlier drafts of the manuscript. Both authors were partially supported by NSF DEB-1455399 to L.K. and Dr. Andrea Wolfe.

- 714 Allman, E. S., Degnan, J. H., and Rhodes, J. A. (2011a). Determining species tree topologies from clade
715 probabilities under the coalescent. *J. Theoret. Biol.*, **289**, 96–106.
- 716 Allman, E. S., Degnan, J. H., and Rhodes, J. A. (2011b). Identifying the rooted species tree from the
717 distribution of unrooted gene trees under the coalescent. *J. Math. Biol.*, **62**, 833–862.
- 718 Allman, E. S., Degnan, J. H., and Rhodes, J. A. (2013). Species tree inference by the star method and its
719 generalizations. *Journal of Computational Biology*, **20**(1), 50–61.
- 720 Andersen, L. N., Mailund, T., and Hobolth, A. (2014). Efficient computation in the im model. *J. Math.*
721 *Biol.*, **68**(6), 1423–1451.
- 722 Bayzid, M., Mirarab, S., Warnow, T., *et al.* (2015). Weighted statistic binning: enabling statistically
723 consistent genome-scale phylogenetic analyses. *PLoS ONE*, *10*(6), e0129183.
- 724 Burgess, R. and Yang, Z. (2008). Estimation of hominoid ancestral population sizes under bayesian coalescent
725 models incorporating mutation rate variation and sequencing errors. *Molecular Biology and Evolution*,
726 **25**(9), 1979–1994.
- 727 DeGiorgio, M. and Degnan, J. H. (2010). Fast and consistent estimation of species trees using supermatrix
728 rooted triples. *Molecular Biology and Evolution*, **27**(3), 552–569.
- 729 Degnan, J. and Salter, L. (2005). Gene tree distributions under the coalescent process. *Evolution*, **59**, 24–37.
- 730 Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees.
731 *PLoS Genetics*, **3**, 762–768.
- 732 Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multi-
733 species coalescent. *Trends Ecol. Evol.*, **24**(6), 332–340.
- 734 Degnan, J. H., DeGiorgio, M., Bryant, D., and Rosenberg, N. A. (2009). Properties of consensus methods
735 for inferring species trees from gene trees. *Syst. Biol.*
- 736 Degnan, J. H., Rosenberg, N. A., and Stadler, T. (2012). The probability distribution of ranked gene trees
737 on a species tree. *Mathematical Biosciences*, **235**(1), 45–55.
- 738 Eckert, A. J. and Carstens, B. C. (2008). Does gene flow destroy phylogenetic signal? the performance of
739 three methods for estimating species phylogenies in the presence of gene flow. *Molecular Phylogenetics*
740 *and Evolution*, **49**(3), 832–842.
- 741 Ewing, G. B., Ebersberger, I., Schmidt, H. A., and Von Haeseler, A. (2008). Rooted triple consensus and
742 anomalous gene trees. *BMC Evolutionary Biology*, **8**(1), 118.
- 743 Fan, H. H. and Kubatko, L. S. (2011). Estimating species trees using approximate bayesian computation.
744 *Molecular Phylogenetics and Evolution*, **59**(2), 354–363.
- 745 Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X.,
746 Hall, A. B., Catteruccia, F., Kakani, E., *et al.* (2015). Extensive introgression in a malaria vector species
747 complex revealed by phylogenomics. *Science*, **347**(6217), 1258524.
- 748 Gerard, D., Gibbs, H. L., and Kubatko, L. (2011). Estimating hybridization in the presence of coalescence
749 using phylogenetic intraspecific sampling. *BMC Evolutionary Biology*, **11**(1), 291.
- 750 Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint
751 demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet*,
752 **5**(10), e1000695.
- 753 Hey, J. (2010). Isolation with migration models for more than two populations. *Molecular Biology and*
754 *Evolution*, **27**(4), 905–920.

- Hey, J. and Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *drosophila pseudoobscura* and *d. persimilis*. *Genetics*, **167**, 747–760.
- Hobolth, A., Andersen, L. N., and Mailund, T. (2011). On computing the coalescence time density in an isolation-with-migration model with few samples. *Genetics*, **187**(4), 1241–1243.
- Huang, H. and Knowles, L. (2009). What’s the biological reality of the anomaly zone? *Syst. Biol.*, **58**, 527–536.
- Huang, H., He, Q., Kubatko, L., and Knowles, L. (2010). Sources of error for species-tree estimation: Impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Syst. Biol.*, **59**, 573–583.
- Huang, H., Tran, L. A., and Knowles, L. L. (2014). Do estimated and actual species phylogenies match? evaluation of east african cichlid radiations. *Molecular Phylogenetics and Evolution*, **78**, 56–65.
- Hudson, R. R. (2002). Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, **18**(2), 337–338.
- Jukes, T. and Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro, editor, *Mammalian protein metabolism*, pages 21–123. Academic Press, New York.
- Kamm, J. A., Terhorst, J., and Song, Y. S. (2015). Efficient computation of the joint sample frequency spectra for multiple populations. *arXiv preprint arXiv:1503.01133*.
- Knowles, L. L., Lanier, H. C., Klimov, P. B., and He, Q. (2012). Full modeling versus summarizing gene-tree uncertainty: method choice and species-tree accuracy. *Molecular phylogenetics and evolution*, **65**(2), 501–509.
- Kubatko, L. S. (2009). Identifying hybridization events in the presence of coalescence via model selection. *Syst. Biol.*, **58**(5), 478–488.
- Lanier, H. C., Huang, H., and Knowles, L. L. (2014). How low can you go? the effects of mutation rate on the accuracy of species-tree estimation. *Molecular phylogenetics and evolution*, **70**, 112–119.
- Leaché, A. D., Harris, R. B., Maliska, M. E., and Linkem, C. W. (2013). Comparative species divergence across eight triplets of spiny lizards (sceloporus) using genomic sequence data. *Genome biology and evolution*, **5**(12), 2410–2419.
- Liu, L. and Edwards, S. V. (2009). Phylogenetic analysis in the anomaly zone. *Syst. Biol.*, **58**(4), 452–460.
- Liu, L., Yu, L., and Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, **10**(1), 302.
- Lohse, K. and Frantz, L. A. (2014). Neandertal admixture in eurasia confirmed by maximum-likelihood analysis of three genomes. *Genetics*, **196**(4), 1241–1251.
- Lohse, K., Harrison, R., and Barton, N. H. (2011). A general method for calculating likelihoods under the coalescent process. *Genetics*, **189**(3), 977–987.
- Maddison, W. P. (1997). Gene trees in species trees. *Syst. Biol.*, **46**, 523–536.
- McCormack, J. E., Huang, H., and Knowles, L. L. (2009). Breaking them down and building them up: Evaluating the accuracy of maximum-likelihood estimates of species trees. *Syst. Biol.*
- McCormack, J. E., Heled, J., Delaney, K. S., Peterson, A. T., and Knowles, L. L. (2011). Calibrating divergence times on species trees versus gene trees: implications for speciation history of aphelocoma jays. *Evolution*, **65**(1), 184–202.
- Meng, C. and Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theor. Pop. Biol.*, **75**, 35–45.

- Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). Astral: genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**(17), i541–i548.
- Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, **5**(5), 568–583.
- Poormohammadi, H., Eslahchi, C., and Tusserkani, R. (2014). Tripnet: A method for constructing rooted phylogenetic networks from rooted triplets. *PLoS ONE* **9**(9): e106531.
- Rambaut, A. and Grassly, N. (1997). SeqGen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. in Biosci.*, **13**, 235–238.
- Solís-Lemus, C. and Ané, C. (2015). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *arXiv preprint arXiv:1509.06075*.
- Swofford, D. L. (2003). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates, Sunderland, Massachusetts.
- Takahata, N. (1989). Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics*, **122**, 957–966.
- Than, C. and Nakhleh, L. (2009). Species tree inference by minimizing deep coalescences. *PLoS Computational Biology*, **5**(9), e1000501.
- Wang, Y. and Hey, J. (2010). Estimating divergence parameters with small samples from a large number of loci. *Genetics*, **184**(2), 363–379.
- Wen, D., Yu, Y., and Nakhleh, L. (2016). Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genet*, **12**(5), e1006006.
- Wu, Y. (2012). Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*, **66**(3), 763–775.
- Yu, Y. and Nakhleh, L. (2015). A maximum pseudo-likelihood approach for phylogenetic networks. *BMC genomics*, **16**(Suppl 10), S10.
- Yu, Y., Than, C., Degnan, J. H., and Nakhleh, L. (2011). Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Syst. Biol.*, **60**(2), 138–149.
- Yu, Y., Degnan, J. H., and Nakhleh, L. (2012). The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet*, **8**(4), e1002660–e1002660.
- Yu, Y., Barnett, R. M., and Nakhleh, L. (2013). Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Syst. Biol.*, **62**(5), 738–751.
- Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. (2014). Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, **111**(46), 16448–16453.
- Zhu, T. and Yang, Z. (2012). Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Molecular Biology and Evolution*, **29**(10), 3131–3142.

*Graphical Abstract