Identifiability of speciation times under the multispecies coalescent

Laura Kubatko*

Department of Statistics, Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210

Alexander Leonard

Department of Mathematics, Miami University, Oxford, OH 45056

Julia Chifman

Department of Mathematics and Statistics, American University, Washington, DC 20016

Abstract

The advent of rapid and inexpensive sequencing technologies has necessitated the development of computationally efficient methods for analyzing sequence data for many genes simultaneously in a phylogenetic framework. The coalescent process is the most commonly used model for linking the underlying genealogies of individual genes with the global species-level phylogeny, but inference under the coalescent model is computationally daunting in the typical inference frameworks (e.g., the likelihood and Bayesian frameworks) due to the dimensionality of the space of both gene trees and species trees. Here we consider estimation of the branch lengths in fixed species trees with three or four taxa, and show that these branch lengths are identifiable. We also show that for three and four taxa simple estimators for the branch lengths can be derived based on observed site pattern frequencies. Properties of these estimators, such as their asymptotic variances and large-sample distributions, are examined, and performance of the estimators is assessed using simulation. Finally, we use these estimators to develop a hypothesis test that can be used to delimit species under the coalescent model for three or four putative taxa .

Keywords: multispecies coalescent, speciation time, species delimitation, identifiability

*Corresponding author

Email addresses: `lkubatko@stat.osu.edu` (Laura Kubatko), `leonaraj@miamioh.edu` (Alexander Leonard), `chifman@american.edu` (Julia Chifman)

1. Introduction

The multispecies coalescent (MSC) model is used to describe the process by which nucleotide sequence data arise from a species tree. The model incorporates two distinct processes: the coalescent process [1, 2, 3] is used to generate gene trees given the species tree, and standard nucleotide substitution models [4] are used to generate sequence data along the gene trees for each gene. Over the past 20 years, many methods for estimation of species trees under the MSC have been proposed (see, e.g., Liu et al. (2009) or Kubatko (2019) for reviews of these methods). One reason that so many different methods have been proposed is that standard approaches to estimation, such as those based in the likelihood or Bayesian frameworks, are computationally intensive and often infeasible for data sets of realistic size. Study of properties of the MSC, such as identifiability of parameters, is thus particularly important for developing effective methods for inference.

Figure 1 shows two four-taxon species trees for which the leaves of the trees, labeled a, b, c, and d, represent present-day species for which DNA sequence data from many loci are observed. The shaded internal branches of the tree, labeled P_1, P_2, and P_3, represent hypothetical common ancestral populations for the species at the leaves. The parameters τ_1, τ_2, and τ_3 are the times from the present back to various speciation events, events in which an ancestral population gives rise to two distinct descendants, indicated by dotted lines. From an inference standpoint, we wish to estimate both the tree topology (i.e., the branching pattern) and the speciation times τ_i, $i = 1, 2, 3$. Figure 1 also depicts gene trees as red lines embedded within a given species tree. These trees represent the phylogenetic histories, which we will denote by h, of the lineages sampled from species a, b, c, and d for a particular gene. The parameters t_1, t_2, and t_3 are the times of coalescent events, events in which two lineages first share a common ancestor, measured from the most immediately descendant speciation event. For example, in Figure 1 (left panel) the second coalescent event occurs in population P_3 and it represents the time, denoted by t_2, from the speciation event that resulted in two distinct populations P_1 and P_2 backward until the lineages c and d share a common ancestor. Note that the coalescent event described in the previous sentence could instead have happened in population P_1, resulting in the same gene tree topology but different coalescent times. Thus, to make this distinction clear, we will associate with each gene tree history h a vector of coalescent times $\mathbf{t}_h$ corresponding to specific intervals between speciation events. We denote this by a pair $(h, \mathbf{t}_h)$. We begin by defining the probability model for general species phylogenies with

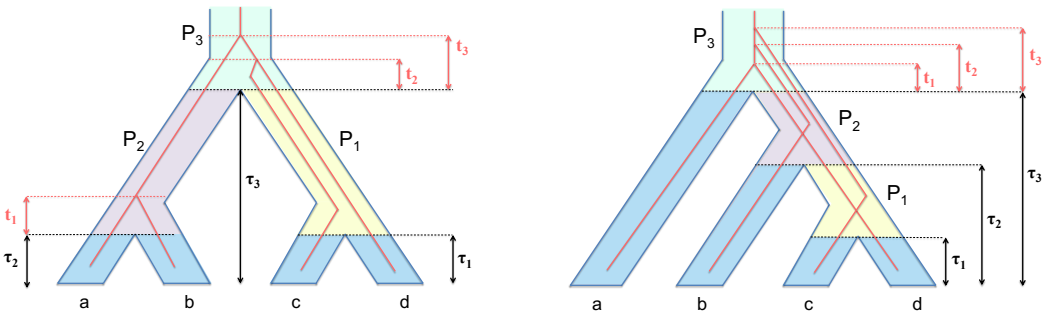

Figure 1: Model symmetric (left) and asymmetric (right) four-taxon species trees with speciation times denoted by τ_1, τ_2, and τ_3. Each speciation time is measured from the present time (the tips of the tree) backward. Note that the definitions of τ_1 and τ_3 are the same in both panels (i.e., τ_1 is the speciation events that separates species c from species d, and τ_3 is the initial speciation event), while the definition of τ_2 differs between the two trees due to their differing shapes. Nested within each of the species trees is an example gene tree history (red lines) that shows the sequence of coalescent events among the sampled lineages for the gene represented. The times of the coalescent events are measured from the immediately descendent speciation event and are denoted by t_1,t_2, and t_3.

any number of tips, before returning to the four-taxon case to establish our main results about species tree branch lengths.

Let M be the number of species under consideration, and label the species $1, 2, \ldots, M$. Suppose that m_j individuals are sampled for species j so that the total number of individuals being analyzed is $\mathcal{M} = \sum_{j=1}^{M} m_j$. Denote the conditional probability density of gene tree history h and associated vector of coalescent times $\mathbf{t}_h$ by $f_{(h,\mathbf{t}_h)|(S,\tau)}$, where the conditioning is on species tree topology S and vector of speciation times τ (a description of how this density is computed can be found in [7]). Define a site pattern as an assignment of states $i_1 i_2 \cdots i_{\mathcal{M}}$ to the $\mathcal{M}$ tips of the tree so that for each $k = 1, 2, \ldots, \mathcal{M}$, $i_k \in \{A, C, G, T\}$. The probability of site pattern $i_1 i_2 \cdots i_{\mathcal{M}}$ arising from *gene tree history* $(h, \mathbf{t}_h)$ is denoted by by $p^h_{(i_1 i_2 \cdots i_{\mathcal{M}})|(h,\mathbf{t}_h)}$. We note that this probability is the typical phylogenetic likelihood for a gene tree, which is computed according to one of the standard nucleotide substitution models. The probability that site pattern $i_1 i_2 \cdots i_{\mathcal{M}}$ arises from the *species tree* is then given by

$$p_{i_1 i_2 \cdots i_{\mathcal{M}}|(S,\tau)} = \sum_{h \in \mathcal{H}} \int_{\mathbf{t}_h} p^h_{(i_1 i_2 \cdots i_n)|(h,\mathbf{t}_h)} f_{(h,\mathbf{t}_h)|(S,\tau)} d\mathbf{t}_h \tag{1}$$

where the sum is over all $\mathcal{H}$ of the possible gene tree histories with the corresponding branch lengths $\mathbf{t}_h$ integrated out. Full details of the calculations are given in [8].

Equation (1) assumes that each site in the sequence alignment is an independent and identically distributed observation from the MSC model. In other words, each site is a draw from the distribution of gene trees given the species tree under the MSC, with mutation along the sampled gene tree then occurring following one of the standard nucleotide substitution models. We refer to data of this type as *coalescent independent sites* (CIS), and contrast this data type with SNP data, which do not include constant sites. With the model specified here, a data set consisting of CIS can be seen as a sample from the multinomial distribution, with $4^{\mathcal{M}}$ possible categories (i.e., the number of possible sites patterns) and with category probabilities given by the site pattern probabilities.

We note that for a fixed species tree, the site pattern probabilities are functions of the species tree branch lengths (the τs in the equation above), as well as the effective population sizes. In the empirical setting in which data are observed and the goal is to estimate the tree topology, branch lengths, and effective population sizes, a necessary first step is to establish that these model parameters are identifiable. Identifiability of the species tree topology (but not the branch lengths) was proved by Chifman and Kubatko (2015) for species trees that satisfy the molecular clock for the GTR+I+Γ substitution model and all submodels thereof. Long and Kubatko (2019) proved identifiability for non-clock species tree topologies and for tree topologies in which the effective population sizes vary throughout the tree. However, identifiability of other model parameters, such as the branch lengths, has not been formally established in any of these cases.

Here, we show that the branch lengths in three- and four-taxon species trees are identifiable under the multispecies coalescent with the JC69 model [10] for the setting considered by [8] in the case in which one lineage is sampled per species and a single effective population size is assumed for the whole tree. The method used to establish indentifiability leads to a natural estimator of the branch lengths for the case of four taxa that can be computed analytically. Variance estimates are also available analytically in this case. In Section 2, we provide a proof of the identifiability of the branch lengths. While our proof uses the site pattern probabilities computed in Chifman and Kubatko (2015), this result was not provided in our previous work. In Section 3, we derive novel formulas for estimators of the branch lengths that are motivated by the proof presented here, along with formulas for the variances of these estimators. Corresponding results for the case of three taxa are given in Appendix C. Section 4 uses simulation to evaluate the performance of these estimators

for simulated data. In Section 5, we describe an application of the branch length estimators to the problem of species delimitation, and we apply this method to an empirical data set in Section 6. We conclude in Section 7 with a discussion of the issues that arise when applying these estimators to larger trees, as well as some applications in which these estimators can be used.

2. Identifiability of the species tree branch lengths

We begin by considering the case in which $M = 4$ and $m_j = 1$ for $j = 1, 2, 3, 4$, i.e., a four-taxon species tree with one lineage sampled per species. In this case, there are 15 possible site patterns given by

$$p_{xxxx},\ p_{xxxy},\ p_{xxyx},\ p_{xyxx},\ p_{yxxx},\ p_{xyxy},\ p_{xyyx},\ p_{xxyy},$$
$$p_{xyxz},\ p_{xyzx},\ p_{yxxz},\ p_{yxzx},\ p_{xxyz},\ p_{yzxx},\ p_{xyzw}$$

where x, y, w, and z denote distinct nucleotides. Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{15})$ denote the vector of counts of these site patterns in an observed data set, and let $\mathbf{p} = (p_1, p_2, \ldots, p_{15})$ denote their probabilities. Then for a sample of n CIS under the assumptions of the MSC, $\mathbf{Y} \sim \text{Multinominal}(n, \mathbf{p})$, and thus we have the following:

$$E(Y_k) = np_k, \quad k = 1, 2, \ldots, 15 \tag{2}$$

$$Var(Y_k) = np_k(1 - p_k), \quad k = 1, 2, \ldots, 15 \tag{3}$$

$$Cov(Y_k, Y_l) = -np_k p_l \quad k, l = 1, 2, \ldots, 15 \tag{4}$$

Under the JC69 model, many of these site patterns will have equal probabilities. For example, Chifman and Kubatko (2015) show that for the rooted symmetric four-taxon species tree, these 15 site pattern probabilities can be further reduced to 9 distinct site pattern probabilities, while for the rooted asymmetric four-taxon species tree, there are 11 distinct site pattern probabilities. Moreover, they give formulas for these site pattern probabilities for both the symmetric and asymmetric four-taxon species trees as a function of the species tree branch lengths (τ) under the assumption of a single effective population size, θ for the entire tree. As is standard in species tree modeling using the coalescent process, we define the effective population size parameter $\theta = 4N_e\mu'$, where N_e is the population size and μ' is mutation rate in units of substitutions per site per generation in the DNA sequence. The branch lengths in the species tree are given in coalescent units, which are the

number of $2N_e$ generations. These site pattern probabilities can be used to establish identifiability, as shown below.

Case 1: The Symmetric Species Tree. Consider first the case of the symmetric species tree (Figure 1). The 15 site pattern probabilities listed above can be reduced to 9 distinct site pattern probabilities in this case:

$$p_{xxxx},\ p_{xxxy} = p_{xxyx},\ p_{xyxx} = p_{yxxx},\ p_{xyxy} = p_{yxxy},\ p_{xxyy},$$

$$p_{xxyz},\ p_{yzxx},\ p_{xyxz} = p_{yxxz} = p_{xyzx} = p_{yxzx},\ p_{xyzw}. \tag{5}$$

Chifman and Kubatko (2015) showed that for $i_k \in \{A, C, G, T\}, k = 1, 2, 3, 4$, these site pattern probabilities can be written as

$$\begin{aligned} p_{i_1 i_2 i_3 i_4|(S,\boldsymbol{\tau})} &= c_0 + c_1 x_1^{2\mu} + c_2 x_2^{2\mu} + c_3 x_1^{2\mu} x_2^{2\mu} + c_4 x_3^{2\mu} + c_5 x_1^{\mu} x_3^{2\mu} + c_6 x_2^{\mu} x_3^{2\mu} \\ &\quad + c_7 x_1^{\mu} x_2^{\mu} x_3^{2\mu} + c_8 x_1^{-\frac{2}{\theta}} x_2^{-\frac{2}{\theta}} x_3^{4(\mu+\frac{1}{\theta})}, \end{aligned}$$

where $x_j = e^{-\tau_j}$ for $j = 1, 2, 3$, $\mu = 4/3$ for the JC69 model (so that gene tree branch lengths are measured as the expected number of mutations per site) , and the coefficients $c_i, i = 0, 1, 2, \ldots, 8$ are given in Appendix A.

Let $(\mathbf{C_s})_{9\times 9}$ be the matrix of coefficients given in Appendix A, and note that the above expressions for the site pattern probabilities can be written in matrix form as

$$\mathbf{C_s}\beta_s = \mathbf{p}_s \tag{6}$$

where

$$\begin{aligned} \beta_s^T &= \left(1, x_1^{2\mu}, x_2^{2\mu}, x_1^{2\mu}x_2^{2\mu}, x_3^{2\mu}, x_1^{\mu}x_3^{2\mu}, x_2^{\mu}x_3^{2\mu}, x_1^{\mu}x_2^{\mu}x_3^{2\mu}, (x_1x_2)^{-2/\theta}x_3^{4(\mu+1/\theta)}\right) \\ &\text{and} \\ \mathbf{p}_s^T &= (p_{xxxx}, p_{xxxy}, p_{xyxx}, p_{xyxy}, p_{xxyy}, p_{xxyz}, p_{yzxx}, p_{xyzx}, p_{xyzw}). \end{aligned}$$

Note that $\mathbf{C_s}$ is full rank, and thus the speciation times τ_1, τ_2, and τ_3 can be identified. In particular, we solve Equation (6) for β_s. Letting $y_i = x_i^{2\mu}$, for $i \in \{1, 2, 3\}$, the unique solution is

$$y_1 = 4(1+\mu\theta)(p_{xxxx} - 2p_{xxxy} + 3p_{xxyy} - 2p_{xxyz} + 6p_{xyxx} - 2p_{xyxy} - 8p_{xyxz} - 2p_{xyzw} + 6p_{yzxx}) \tag{7}$$

$$y_2 = 4(1+\mu\theta)(p_{xxxx} + 6p_{xxxy} + 3p_{xxyy} + 6p_{xxyz} - 2p_{xyxx} - 2p_{xyxy} - 8p_{xyxz} - 2p_{xyzw} - 2p_{yzxx}) \tag{8}$$

$$y_3 = 4(1+\mu\theta)(p_{xxxx} + 2p_{xxxy} - p_{xxyy} - 2p_{xxyz} + 2p_{xyxx} + 2p_{xyxy} - 2p_{xyzw} - 2p_{yzxx}). \tag{9}$$

Recalling that $x_i = e^{-\tau_i}$, we convert back to speciation times using:

$$\tau_i = -\ln(y_i^{\frac{1}{2\mu}}). \tag{10}$$

Case 2: The Asymmetric Species Tree. We now consider the case of the asymmetric species tree (Figure 1). In this case, the 15 possible site pattern probabilities can be reduced to 11 distinct probabilities:

$$p_{xxxx},\ p_{xxxy} = p_{xxyx},\ p_{xyxx},\ p_{yxxx},\ p_{xxyy},\ p_{xyxy} = p_{yxxy},$$

$$p_{xxyz},\ p_{yzxx},\ p_{xyxz} = p_{xyzx},\ p_{yxxz} = p_{yxzx},\ p_{xyzw}.$$

Chifman and Kubatko (2015) showed that for $i_k \in \{A, C, G, T\}, k = 1, 2, 3, 4$, the site pattern probabilities can be written as

$$\begin{aligned} p_{i_1 i_2 i_3 i_4|(S,\boldsymbol{\tau})} &= c_0 + c_1 x_1^{2\mu} + c_2 x_2^{2\mu} + c_3 x_1^{\mu} x_2^{2\mu} + c_4 x_3^{2\mu} + c_5 x_1^{\mu} x_3^{2\mu} + c_6 x_1^{2\mu} x_3^{2\mu} \\ &\quad + c_7 x_2^{\mu} x_3^{2\mu} + c_8 x_1^{\mu} x_2^{\mu} x_3^{2\mu} + c_9 x_1^{\frac{-2}{\theta}} x_2^{2\left(\mu+\frac{1}{\theta}\right)} x_3^{2\mu}, \end{aligned}$$

where $x_j = e^{-\tau_j}$ for $j = 1, 2, 3$, $\mu = 4/3$ for the JC69 model, and the coefficients in the above expression are given in Appendix A. Letting $(\mathbf{C_a})_{11\times 10}$ be the matrix of coefficients, the above expressions for the site pattern probabilities can be written as

$$\mathbf{C_a}\beta_a = \mathbf{p}_a \tag{11}$$

where

$$\beta_a^T = \left(1, x_1^{2\mu}, x_2^{2\mu}, x_1^{\mu}x_2^{2\mu}, x_3^{2\mu}, x_1^{\mu}x_3^{2\mu}, x_1^{2\mu}x_3^{2\mu}, x_2^{\mu}x_3^{2\mu}, x_1^{\mu}x_2^{\mu}x_3^{2\mu}, x_1^{\frac{-2}{\theta}} x_2^{2\left(\mu+\frac{1}{\theta}\right)} x_3^{2\mu}\right)$$

and

$$\mathbf{p}_a^T = \left(p_{xxxx}, p_{xxxy}, p_{xyxx}, p_{yxxx}, p_{xxyy}, p_{xyxy}, p_{xxyz}, p_{yzxx}, p_{xyxz}, p_{yxxz}, p_{xyzw}\right).$$

Since matrix $\mathbf{C_a}$ has full column rank, we can solve Equation (11) for β_a using the left-inverse of $\mathbf{C_a}$ to obtain the following estimators

$$\begin{aligned} y_1 &= \frac{4}{3}(1+\mu\theta)(3p_{xxxx} - 7p_{xxxy} + 8p_{xxyy} - 8p_{xxyz} + 10p_{xyxx} - 5p_{xyxy} - 10p_{xyxz} \\ &\quad - 6p_{xyzw} + 9p_{yxxx} - 12p_{yxxz} + 18p_{yzxx}) \\ y_2 &= \frac{4}{3}(1+\mu\theta)(3p_{xxxx} + 5p_{xxxy} - 4p_{xxyy} - 8p_{xxyz} - 2p_{xyxx} + 7p_{xyxy} - 10p_{xyxz} \\ &\quad - 6p_{xyzw} + 9p_{yxxx} + 12p_{yxxz} - 6p_{yzxx}) \\ y_3 &= \frac{4}{3}(1+\mu\theta)(3p_{xxxx} + 10p_{xxxy} + p_{xxyy} + 2p_{xxyz} + 5p_{xyxx} + 2p_{xyxy} + 4p_{xyxz} \\ &\quad - 6p_{xyzw} - 3p_{yxxx} - 12p_{yxxz} - 6p_{yzxx}), \end{aligned}$$

where $y_i = x_i^{2\mu}$ for $i \in \{1, 2, 3\}$. The following expression for the left inverse was used $(\mathbf{C_a}^T\mathbf{C_a})^{-1}\mathbf{C_a}^T$. We solve for τ_i as before:

$$\tau_i = -\ln(y_i^{\frac{1}{2\mu}})$$

Identifiability of the speciation times. The work above establishes that the species tree branch lengths are identifiable for both the symmetric and asymmetric species trees under the MSC for the JC69 model for a four-taxon species tree. We note that we used the JC69 model for our initial verification of identifiability because closed-form expressions for the site patterns probabilities are available. But we note that for any sub-model of GTR+I+Γ expressions analogous to Equations (6) and (11) can be written once the parameters of the substitution model are specified. However, it is not immediately obvious under what models and conditions the system $\mathbf{p} = \mathbf{C}\beta$ will be consistent, and this will require careful investigation. If the system is consistent, then it is possible that for some sub-models the matrix $\mathbf{C}$ will have full column rank, in which case the left inverse exists and a unique solution can be computed. It is also possible that the system $\mathbf{p} = \mathbf{C}\beta$ will be underdetermined for some sub-models but still consistent. In this case there are infinitely many solutions and the pseudoinverse can be used to find the minimum length solution of the system. If there are parameter choices for which the system is inconsistent, the pseudoinverse can be used again and, in this case, provides the least-squares solution of minimum length.

3. Moment-based estimators of the branch lengths

The previous section provides explicit analytic formulas for the branch lengths for both the symmetric and asymmetric rooted four-taxon species trees under the MSC for the JC69 model. In this section, we show how the variances of these estimators can be estimated.

To illustrate the main ideas, consider the case of the symmetric species tree in Figure 1 and suppose that our interest is in estimating $\tau_1 = -\ln(y_1^{1/2\mu})$. Let $\mathbf{Y_s} = (Y_{s,1}, Y_{s,2}, \ldots, Y_{s,9})$ be the observed counts of the site patterns corresponding to each of the patterns listed in (5) above, and recall that under the MSC model for a sample of n CIS, $\mathbf{Y} \sim \text{Multinomial}(n, \mathbf{p}_s)$. Thus, for large n, we have

$$\hat{p}_{s,k} = \frac{Y_{s,k}}{n} \sim N\left(p_{s,k}, \frac{p_{s,k}(1 - p_{s,k})}{n}\right) \tag{12}$$

$$cov(\hat{p}_{s,k}, \hat{p}_{s,l}) = \frac{-p_{s,k}p_{s,l}}{n} \tag{13}$$

for $k, l = 1, 2, \ldots, 9$.

To estimate $y_1 = x_1^{2\mu}$, we can substitute the empirical estimates $\hat{\mathbf{p}}_s$ into Equation (7) for $\mathbf{p}_s$. We denote this estimator $\hat{y}_1$. We can also find an expression for the asymptotic variance of this estimate by taking the variance in Equation (7) and substituting the appropriate variance and covariance terms from (12) and (13) to get

$$\begin{aligned} Var(\hat{y}_1) = 16(1+\mu\theta)^2(&\sigma_1 + 4\sigma_2 + 9\sigma_5 + 4\sigma_6 + 36\sigma_3 + 4\sigma_4 + 64\sigma_8 + 4\sigma_9 + 36\sigma_7 \\ &- 4\sigma_{1,2} + 6\sigma_{1,5} - 4\sigma_{1,6} + 12\sigma_{1,3} - 4\sigma_{1,4} - 16\sigma_{1,8} - 4\sigma_{1,9} + 12\sigma_{1,7} - 12\sigma_{2,5} + 8\sigma_{2,6} \\ &- 24\sigma_{2,3} + 8\sigma_{2,4} + 32\sigma_{2,8} + 8\sigma_{2,9} - 24\sigma_{2,7} - 12\sigma_{5,6} + 36\sigma_{3,5} - 12\sigma_{4,5} - 48\sigma_{5,8} - 12\sigma_{5,9} \\ &+ 36\sigma_{5,7} - 24\sigma_{3,6} + 8\sigma_{4,6} + 32\sigma_{6,8} + 8\sigma_{6,9} - 24\sigma_{6,7} - 24\sigma_{3,4} - 96\sigma_{3,8} - 24\sigma_{3,9} + 72\sigma_{3,7} \\ &+ 32\sigma_{4,8} + 8\sigma_{4,9} - 24\sigma_{4,7} + 32\sigma_{8,9} - 96\sigma_{7,8} - 24\sigma_{7,9}) \end{aligned} \tag{14}$$

where $\sigma_i = var(\hat{p}_i)$ and $\sigma_{i,j} = cov(\hat{p}_i, \hat{p}_j)$. We estimate all of the quantities in the expression above using the empirical frequencies to obtain $\widehat{Var}(\hat{y}_1)$, which is an estimator of this asymptotic variance.

Similar expressions can be derived for all other branch lengths in the both the symmetric and the asymmetric species tree. We provide these expressions in Appendix B. Note that it is easy to convert $\hat{y}_i$ to the original scale of τ, using

$$\hat{\tau}_i = -\ln\left(\hat{y}_i^{1/2\mu}\right). \tag{15}$$

When n is large, $\hat{y}_i$ will be approximately normally distributed with mean y_i and variance that can be approximated by $\widehat{Var}(\hat{y}_i)$. Furthermore, the asymptotic normal distribution can be used to construct a $(1-\alpha) \times 100\%$ confidence interval using

$$\hat{y}_i \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{y}_i)} \tag{16}$$

where $z_{\alpha/2}$ is the value such that $\Phi(z_{\alpha/2}) = 1-\alpha/2$ where Φ is the cumulative distribution function of the standard normal distribution. This interval can be transformed to the original scale of τ by applying Equation (15) to both endpoints.

To summarize, the estimators derived here can be used to obtain unbiased point estimates of the speciation times, as well as corresponding large-sample confidence intervals. It is to be noted that all of this can be achieved with closed-form equations, i.e., no optimization is required, and thus these quantities can be computed instantaneously. In the next section, we use simulation to show that these estimators behave as expected.

4. Simulations

We examine the performance of the estimators using simulation. The overall simulation design is as follows:

1. Simulate n CIS from Multinomial(n,$\mathbf{p_s}$) or Multinomial(n,$\mathbf{p_a}$).
2. Compute $\hat{y}_i$ and $\widehat{Var}(\hat{y}_i)$ for $i = 1, 2, 3$.
3. Construct 95% confidence intervals for τ_i, $i = 1, 2, 3$ by transforming the endpoints of the interval in Equation (16) using (15).
4. Repeat steps 1-3 B times.

In Step 1, we must specify values for all of the parameters. We consider the settings below for both the symmetric and the asymmetric model trees in Figure 1. The branch lengths $\tau_i, i = 1, 2, 3$ are given in coalescent units (number of $2N_e$ generations) and θ is the effective population size.

Setting 1: $\tau_1 = 1.0$, $\tau_2 = 1.5$, $\tau_3 = 2.5$, $\theta = 0.001$, $n = 100,000$
Setting 2: $\tau_1 = 1.0$, $\tau_2 = 1.5$, $\tau_3 = 2.5$, $\theta = 0.005$, $n = 100,000$
Setting 3: $\tau_1 = 1.0$, $\tau_2 = 1.5$, $\tau_3 = 2.0$, $\theta = 0.001$, $n = 100,000$
Setting 4: $\tau_1 = 1.0$, $\tau_2 = 1.5$, $\tau_3 = 2.0$, $\theta = 0.005$, $n = 100,000$
Setting 5: $\tau_1 = 1.0$, $\tau_2 = 1.5$, $\tau_3 = 2.5$, $\theta = 0.001$, $n = 10,000$
Setting 6: $\tau_1 = 1.0$, $\tau_2 = 1.5$, $\tau_3 = 2.5$, $\theta = 0.005$, $n = 10,000$
Setting 7: $\tau_1 = 1.0$, $\tau_2 = 1.5$, $\tau_3 = 2.0$, $\theta = 0.001$, $n = 10,000$
Setting 8: $\tau_1 = 1.0$, $\tau_2 = 1.5$, $\tau_3 = 2.0$, $\theta = 0.005$, $n = 10,000$

Settings 1-4 consider a total of $n = 100,000$ CIS, which is a moderately large dataset size that is typical of many genome-scale phylogenomics studies. Settings 5-8 consider a smaller dataset size, with only $n = 10,000$ CIS. In odd numbered settings, the effective population size parameter θ is set to 0.001, while in the even numbered settings, $\theta = 0.005$. These represent typical values that are on the lower (odd settings) and upper (even settings) range of values commonly observed. The values of τ_1 and τ_2 are held constant in all of the settings at values that are typical but challenging, since they represent relatively short branch lengths. The value of τ_3 is varied to provide both a very difficult problem ($\tau_3 = 2.0$, creating an internal branch of length 0.5 coalescent units; settings 3, 4, 7, and 8) or a slightly easier problem ($\tau = 2.5$, creating a branch of length 1.0 coalescent units; settings 1, 2, 5, and 6). For all simulation settings, $B = 25,000$.

For each simulation setting, histograms of the estimates $\hat{y}_i$ and $\widehat{Var}(\hat{y}_i)$ are constructed (Figures 2 and 3 show these histograms for Setting 1; figures for all other settings are shown in the Supplemental Material). The thick gray line corresponds to the value of the parameter estimated across the B simulation replicates (i.e., for the histogram of the estimates $\hat{y}_i$, the gray line corresponds to the mean of the estimates across the B replicates; for the histogram of the estimates $\widehat{Var}(\hat{y}_i)$, the gray line corresponds to the sample variances of the estimates $\hat{y}_i$ across the B replicates). The blue lines in each plot give the "true" values of the parameters, i.e., the values computed by plugging the true site pattern probabilities into the expressions for the estimators. We note that, in general, it is easier to estimate a mean than a variance and thus the blue lines and gray lines agree more closely for the plots of $\hat{y}_i$. Additionally, we plot 95% confidence intervals for $\tau_i, i = 1, 2, 3$, for the first 100 replicates, with a red line corresponding to the true value of τ_i. The observed portion of the B replicates that contain the true value is given in the plot. We show results for Setting 1 here for both the symmetric (Figure 2) and asymmetric (Figure 3) species trees. The Supplemental Material contains the plots for the other simulation settings. In addition, an R Markdown file that will allow the reader to reproduce these simulation results as well as examine different settings is available at https://github.com/lkubatko/MOMSpeciationTimes.

An examination of subfigures (a) - (c) in Figures 2 and 3 indicates that the sampling distribution for $\hat{y}_i$, $i = 1, 2, 3$, appears to be normally distributed, as expected, and that y_i is unbiasedly estimated. Subfigures (d) - (f) in each panel show the distribution of estimated variances, again indicating that the variance is accurately estimated. We note that the value estimated from the simulated values of $\hat{y}_i$, represented by the gray line in the figures, is further from both the simulated estimates of the variance and the true value in the case of asymmetric trees. This is because a larger number of simulation replicates is needed to accurately estimate variances than means and because the variances are quite small in this setting. Because the variance for the asymmetric tree contains more terms than does the variance for the symmetric tree (see the expressions in Appendix B), the variance is more difficult to estimate by using the variance of the estimates in this case. However, the accuracy of our estimated variances is supported by the strong performance of our confidence intervals for the speciation times. In particular, we note that all 95% confidence intervals show coverage probabilities very close to 95% – across all 16 simulation settings and all three parameters, the empirical coverage was always between 93.7% and 96.0%, suggesting that

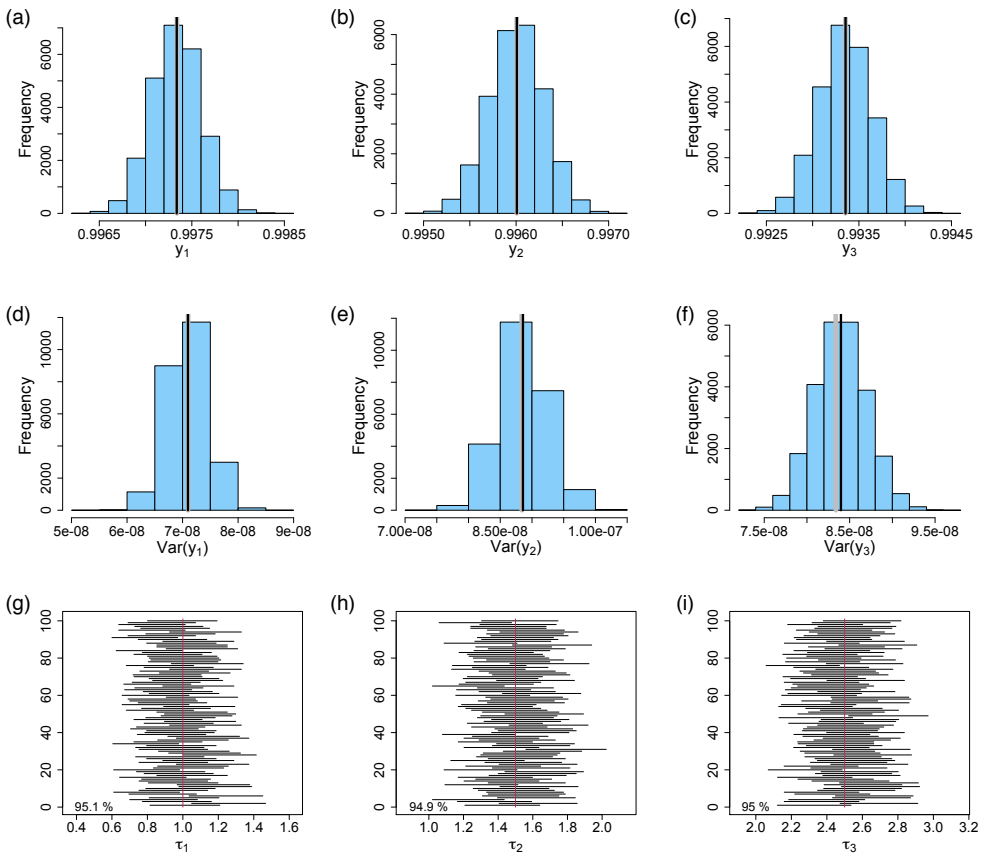

Figure 2: Results of the simulation for Setting 1 for the symmetric species tree in Figure 1. The first row shows histograms of (a) $\hat{x}_1^{2\mu}$, (b) $\hat{x}_2^{2\mu}$, and (c) $\hat{x}_3^{2\mu}$. The second row shows histograms of (d) $\widehat{Var}(\hat{x}_1^{2\mu})$, (e) $\widehat{Var}(\hat{x}_2^{2\mu})$, and (f) $\widehat{Var}(\hat{x}_3^{2\mu})$. In the first two rows, the gray line corresponds to the value obtained by summarizing the simulated values, and the blue lines corresponds to the true values (see text for details). The third row plots the first 100 of the simulated 95% confidence intervals for the (g) τ_1, (h) τ_2, and (i) τ_3. The red line indicates the true value, and the number in the bottom left corner is the portion of all simulation replicates for which the confidence interval includes the true value.

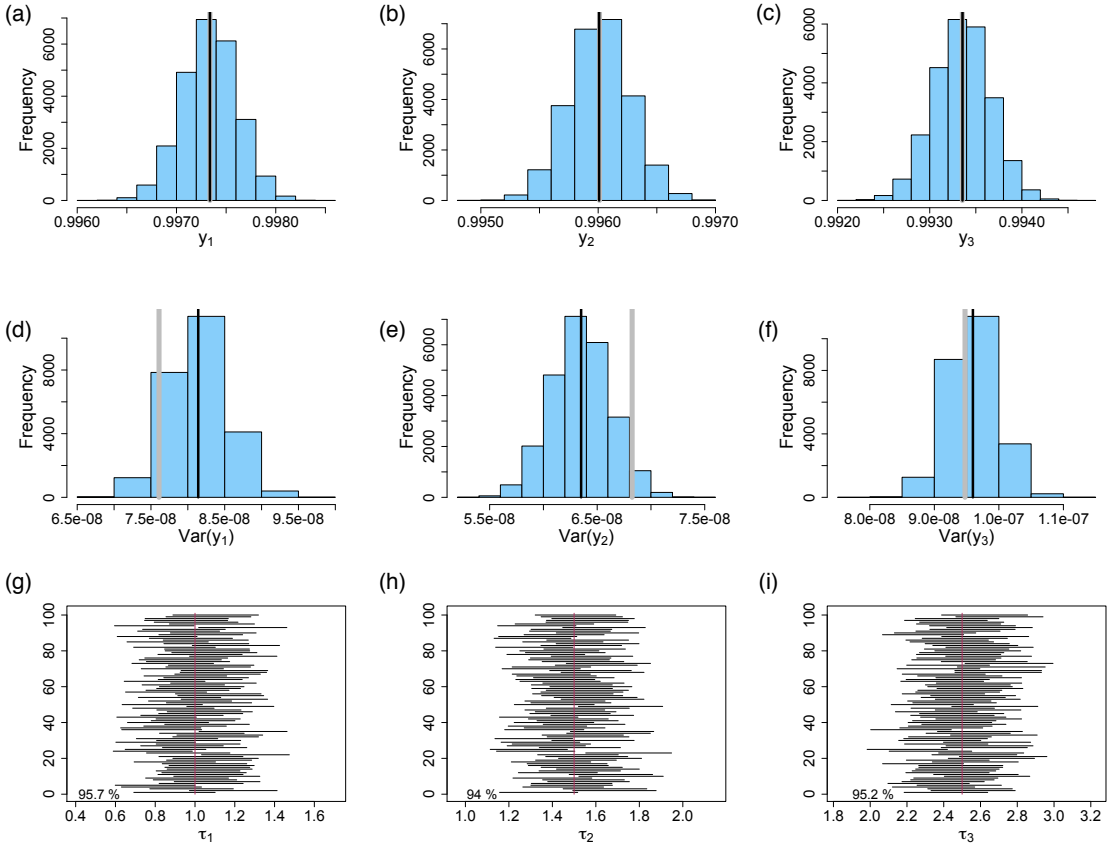

Figure 3: Results of the simulation for Setting 1 for the asymmetric species tree in Figure 1. The first row shows histograms of (a) $\hat{x}_1^{2\mu}$, (b) $\hat{x}_2^{2\mu}$, and (c) $\hat{x}_3^{2\mu}$. The second row shows histograms of (d) $\widehat{Var}(\hat{x}_1^{2\mu})$, (e) $\widehat{Var}(\hat{x}_2^{2\mu})$, and (f) $\widehat{Var}(\hat{x}_3^{2\mu})$. In the first two rows, the gray line corresponds to the value obtained by summarizing the simulated values, and the blue lines corresponds to the true values (see text for details). The third row plots the first 100 of the simulated 95% confidence intervals for the (g) τ_1, (h) τ_2, and (i) τ_3. The red line indicates the true value, and the number in the bottom left corner is the portion of all simulation replicates for which the confidence interval includes the true value.

our estimates of both the branch lengths and their variances are very accurate.

5. Application to species delimitation

The estimators derived in the previous section can be applied to the problem of species delimitation, in which one seeks to determine whether two sets of taxa are members of distinct species (vs. members of the same species). For example, consider the symmetric species tree in Figure 1, and suppose that we are interested in addressing the question of whether lineages sampled from tip c form a distinct species from those sampled from tip d. One way to ask this question is to ask whether the speciation time τ_1 is significantly larger than 0, since if c and d are not distinct species, then it should be the case that $\tau_1 = 0$.

Formally, we propose to test the hypotheses

$$H_0 : \tau_1 = 0 \text{ vs. } H_1 : \tau_1 > 0$$

using the test statistic

$$Z = \frac{\hat{y}_1 - 1}{\sqrt{\widehat{Var}(\hat{y}_1)}} \tag{17}$$

where we subtract one in the numerator because $\tau_1 = 0$ implies $y_1 = 1$. When the number of sites, n, is large, $Z \sim N(0, 1)$ when H_0 is true, and we reject H_0 when $Z > z_\alpha$.

We test the performance of this method using simulation. For Settings 1, 2, 5, and 6, we generated $B = 1,000$ data sets for $\tau_1 \in \{0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, and computed the test statistic in expression (17) above. For each data set, we computed the power of the test, the proportion of times that the test rejected H_0 at level $\alpha = 0.05$. The power as a function of τ_1 is plotted in Figure 4. To verify the normal approximation, we also plot the test statistics simulated for three choices of τ_1 ($\tau_1 = 0, 0.1, 1.0$) in Figure S15 in the Supplemental Material.

It is easy to see that the test performs well. Figure S15 shows that the test statistics are well-approximated by the normal distribution. The power of the test (Figure 4) increases with increasing τ_1, with the test showing power near 100% for $\tau_1 > 0.1$ when $\theta = 0.005$ and for $\tau_1 > 0.4$ when $\theta = 0.001$. The level of the test (the power computed at the null hypothesis value of $\tau_1 = 0$) is also reasonable (0.049 and 0.044 when $\theta = 0.001$ for the symmetric and asymmetric trees, respectively,

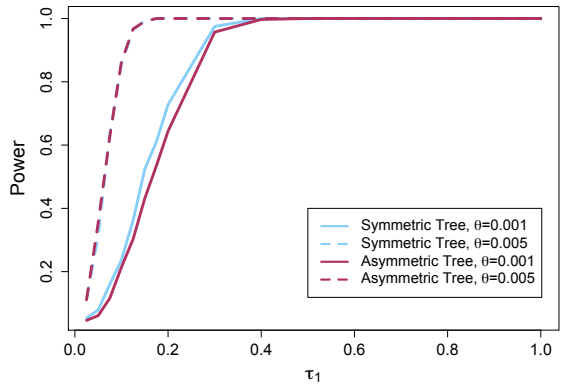

Figure 4: Power of the hypothesis test for both the symmetric (blue) and the asymmetric (red) trees for Setting 1 ($\theta = 0.001$, solid lines) and Setting 2 ($\theta = 0.005$, dashed lines).

and 0.112 and 0.107 when $\theta = 0.005$ for the symmetric and asymmetric trees, respectively), close to the expected value of 0.05 in all cases.

6. Empirical Example: *Sistrurus* rattlesnakes

We demonstrate the application of our methodology to an example data set for *Sistrurus* rattlesnakes. Kubatko et al. (2011) analyzed sequence data collected from 19 genes for two species of rattlesnakes, *Sistrurus catenatus* and *S. miliarius*, each of which consists of 3 identified subspecies. Among the questions addressed by these authors was whether one of the subspecies of *S. catenatus* had diverged sufficiently from the other two that it could be considered a distinct species. Kubatko et al. (2011) provided support for this new species designation based on both genetic analyses and morphological distinctiveness. We use the method described in the previous section to assess support for this conclusion.

Referring to the symmetric tree in Figure 1 (left), we consider data for the following subspecies: a = *S. m. miliarius*; b = *S. m. streckeri*; c = *S. c. catenatus*; and d = *S. c. tergeminus*. To test whether our data support *S. c. catenatus* as a distinct species, we test the null hypothesis that $\tau_1 = 0$ using the test statistic in Equation (17). Selecting one individual of each of the species listed above gives the vector of observed site patterns $(7909, 127, 46, 0, 103, 9, 1, 0, 0)$, ordered as in (5). We assume that $\theta = 0.002$ and $\mu = 4.0/3.0$, corresponding to the Jukes-Cantor model. Using expressions (7), (14), (15), and (16), a 95% confidence interval for τ_1 is $(2.49, 3.91)$. The

test statistic for the hypothesis test is $z = -8.94$, resulting in a p-value of 1.94×10^{-19}. Thus, we reject the null hypothesis that $\tau_1 = 0$, meaning that our analysis shows support for *S. c. catenatus* as a distinct species from *S. c. tergeminus* in agreement with the findings of [11].

We also consider estimation of τ_2 and find an estimate of $\hat{\tau}_2 = 0.44$ with 95% confidence interval $(0.03, 0.85)$, suggesting a much more recent divergence between subspecies within *S. miliarius*, again consistent with [11]. Our estimates are given here in coalescent units, while [11] reported estimates in millions of years before the present. However, comparing our estimates with theirs (see their Table 5), we find similar results: within *S. catenatus*, a large divergence time is inferred for the separation of subspecies *S. c. catenatus* from *S. c. tergeminus*, while within species *S. miliarius*, the divergence among subspecies is inferred to be much more recent. We note that the analysis in [11] was carried out with a computationally-intensive Bayesian approach requiring hours of computing time, while our analysis can be done with a standard calculator.

7. Discussion and conclusions

While our earlier work [8] established identifiability of the species tree topology under the MSC, the work presented here is the first to establish identifiability of the speciation times (or, equivalently, the branch lengths) in a species-level phylogenetic tree under the multispecies coalescent model. In particular, we show that for both the symmetric and asymmetric four-taxon species trees that satisfy the molecular clock, the speciation times (the times of the internal nodes) can be identified for the standard nucleotide substitution models. For the JC69 model [10], we provide specific formulas for estimating these branch lengths, and show that they display the expected large-sample behavior, i.e., they are unbiased and approximately normally distributed. Moreover, simple formulas for estimating the variances of these estimators can be derived. Using properties of the normal distribution, we construct confidence intervals for the speciation times based on these estimators, and use simulation to show that these achieve the expected coverage. We also develop a hypothesis test for speciation that can be used to aid in species delimitation, and we use simulation to show that the test has good power to detect speciation while still achieving the nominal significance level under the null hypothesis.

All of the work here has assumed that the effective population size parameter, θ, is known. In fact, the solutions to Equations (6) and (11) provide relationships that this parameter must satisfy

and can be used to estimate θ, as well, in the case of the JC69 model. However, there is not a closed form solution for θ and its solution must be iterated with the values of τ. Another assumption in this work is that the species tree satisfies the molecular clock. While this assumption is quite commonly made when the goal is to estimate a species-level phylogeny, it is interesting to consider the case when the clock does not hold. The work of Long and Kubatko (2019) provides site pattern probabilities in the non-clock case, and could be used to examine questions of identifiability in this setting.

Finally, we note that our work here has examined only four-taxon species trees. While one might hope to scale this approach up to larger trees, especially because the estimators derived here consist of formulas that can be rapidly computed, this approach is complicated by the fact that expressions for the site pattern probabilities for trees containing more than four species have not been derived, owing to the computational complexity of this setting (see, e.g., Chifman and Kubatko (2015) or Kubatko (2019) for details). Nonetheless, the site pattern probabilities and multinomial distribution function used here have led to estimation of the speciation times for larger trees using a composite likelihood approach [12]. However, this approach requires optimization of the composite likelihood function, and thus is less computationally efficient than the method presented here. Therefore, the method here is useful for addressing questions about speciation times and species designations when the primary interest is in examining a small collection of closely related taxa.

8. Acknowledgements

We thank Elizabeth Allman, John Rhodes, and Seth Sullivant for helpful discussions on this work. L.K. and J.C. were supported by funding from the National Science Foundation (DMS 11-06706).

References

[1] J. F. C. Kingman, On the genealogy of large populations, J. Appl. Prob. 19A (1982) 27–43.

[2] J. F. C. Kingman, The coalescent, Stoch. Proc. Appl. 13 (1982) 235–248.

[3] J. F. C. Kingman, Exchangeability and the evolution of large populations. Pp. 97-112 *in* G. Koch and F. Spizzichino, eds. Exchangeability in probability and statistics. North-Holland: Amsterdam. (1982).

[4] P. Liò, N. Goldman, Models of molecular evolution and phylogeny, Genome Research 8 (1998) 1233–1244.

[5] L. Liu, L. Yu, L. Kubatko, D. K. Pearl, S. V. Edwards, Coalescent methods for estimating multilocus phylogenetic trees, Molecular Phylogenetics and Evolution 53 (2009) 320–328.

[6] L. Kubatko, The multispecies coalescent, in: D. J. Balding, I. Moltke, J. Marioni (Eds.), Handbook of Statistical Genetics, 4th Edition, Wiley, 2019, pp. 219–246.

[7] B. Rannala, Z. Yang, Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci, Genetics 164 (2003) 1645–1656.

[8] J. Chifman, L. Kubatko, Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites, Journal of Theoretical Biology 374 (2015) 35–47.

[9] C. Long, L. Kubatko, Identifiability and reconstructibility of species phylogenies under a modified coalescent, Bulletin of Mathematical Biology 81 (2019) 408–430.

[10] T. Jukes, C. R. Cantor, Evolution of protein molecules, in: H. N. Munro (Ed.), Mammalian protein metabolism, Academic Press, New York, 1969, pp. 21–123.

[11] L. S. Kubatko, H. L. Gibbs, E. Bloomquist, Inferring species-level phylogenies using multi-locus data for a recent radiation of Sistrurus rattlesnakes, Syst. Biol. 60 (4) (2011) 393–409.

[12] J. Peng, D. Swofford, L. Kubatko, Estimation of speciation times under the multispecies coalescent, available on bioRxiv: https://www.biorxiv.org/content/10.1101/681023v1 (2020).

[13] T. Zhu, Z. Yang, Complexity of the simplest species tree problem, Molecular Biology and Evolution 39 (9) (2021) 3993–4009.

Appendix A.

Table 1: Coefficients for site pattern probabilities: symmetric species 4-leaf tree

	xxxx	xxxy	xyxx
c_0	$\frac{1}{256}$	$\frac{1}{256}$	$\frac{1}{256}$
c_1	$\frac{3}{256(1+\mu\theta)}$	$-\frac{1}{256(1+\mu\theta)}$	$\frac{3}{256(1+\mu\theta)}$
c_2	$\frac{3}{256(1+\mu\theta)}$	$\frac{3}{256(1+\mu\theta)}$	$-\frac{1}{256(1+\mu\theta)}$
c_3	$\frac{9}{256(1+\mu\theta)^2}$	$-\frac{3}{256(1+\mu\theta)^2}$	$-\frac{3}{256(1+\mu\theta)^2}$
c_4	$\frac{12}{256(1+\mu\theta)}$	$\frac{4}{256(1+\mu\theta)}$	$\frac{4}{256(1+\mu\theta)}$
c_5	$\frac{24}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$	$\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$
c_6	$\frac{24}{256(1+\mu\theta)(2+\mu\theta)}$	$\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$
c_7	$\frac{48}{256(1+\mu\theta)(2+\mu\theta)^2}$	$-\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$	$-\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$
c_8	$\frac{6\mu\theta(4+\mu\theta)(4+3\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$-\frac{2\mu\theta(4+\mu\theta)(4+3\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$-\frac{2\mu\theta(4+\mu\theta)(4+3\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$

	xyxy	xxyy	xxyz
c_0	$\frac{1}{256}$	$\frac{1}{256}$	$\frac{1}{256}$
c_1	$-\frac{1}{256(1+\mu\theta)}$	$\frac{3}{256(1+\mu\theta)}$	$-\frac{1}{256(1+\mu\theta)}$
c_2	$-\frac{1}{256(1+\mu\theta)}$	$\frac{3}{256(1+\mu\theta)}$	$\frac{3}{256(1+\mu\theta)}$
c_3	$\frac{1}{256(1+\mu\theta)^2}$	$\frac{9}{256(1+\mu\theta)^2}$	$-\frac{3}{256(1+\mu\theta)^2}$
c_4	$\frac{4}{256(1+\mu\theta)}$	$-\frac{4}{256(1+\mu\theta)}$	$-\frac{4}{256(1+\mu\theta)}$
c_5	$-\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$	$\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$
c_6	$-\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$
c_7	$\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$	$-\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$	$\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$
c_8	$\frac{\mu\theta(32+40\mu\theta+10\mu^2\theta^2)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$\frac{2\mu\theta(4+\mu\theta)^2}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$\frac{2\mu^2\theta^2(4+\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$

	yzxx	xyzx	xyzw
c_0	$\frac{1}{256}$	$\frac{1}{256}$	$\frac{1}{256}$
c_1	$\frac{3}{256(1+\mu\theta)}$	$-\frac{1}{256(1+\mu\theta)}$	$-\frac{1}{256(1+\mu\theta)}$
c_2	$-\frac{1}{256(1+\mu\theta)}$	$-\frac{1}{256(1+\mu\theta)}$	$-\frac{1}{256(1+\mu\theta)}$
c_3	$-\frac{3}{256(1+\mu\theta)^2}$	$\frac{1}{256(1+\mu\theta)^2}$	$\frac{1}{256(1+\mu\theta)^2}$
c_4	$-\frac{4}{256(1+\mu\theta)}$	0	$-\frac{4}{256(1+\mu\theta)}$
c_5	$-\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	0	$\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$
c_6	$\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$	0	$\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$
c_7	$\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$	0	$-\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$
c_8	$\frac{2\mu^2\theta^2(4+\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$-\frac{\mu^2\theta^2(4+2\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$\frac{2\mu^3\theta^3}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$

Table 2: Coefficients for site pattern probabilities: asymmetric species 4-leaf tree

	xxxx	xxxy	xyxx	yxxx
c_0	$\frac{1}{256}$	$\frac{1}{256}$	$\frac{1}{256}$	$\frac{1}{256}$
c_1	$\frac{3}{256(1+\mu\theta)}$	$-\frac{1}{256(1+\mu\theta)}$	$\frac{3}{256(1+\mu\theta)}$	$\frac{3}{256(1+\mu\theta)}$
c_2	$\frac{6}{256(1+\mu\theta)}$	$\frac{2}{256(1+\mu\theta)}$	$-\frac{2}{256(1+\mu\theta)}$	$\frac{6}{256(1+\mu\theta)}$
c_3	$\frac{12}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	$\frac{12}{256(1+\mu\theta)(2+\mu\theta)}$
c_4	$\frac{9}{256(1+\mu\theta)}$	$\frac{5}{256(1+\mu\theta)}$	$\frac{5}{256(1+\mu\theta)}$	$-\frac{3}{256(1+\mu\theta)}$
c_5	$\frac{6}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	$\frac{12}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$
c_6	$\frac{9}{256(1+\mu\theta)^2}$	$-\frac{3}{256(1+\mu\theta)^2}$	$-\frac{3}{256(1+\mu\theta)^2}$	$-\frac{3}{256(1+\mu\theta)^2}$
c_7	$\frac{24}{256(1+\mu\theta)(2+\mu\theta)}$	$\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$
c_8	$\frac{48}{256(1+\mu\theta)(2+\mu\theta)^2}$	$-\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$	$-\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$	$-\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$
c_9	$\frac{6\mu\theta(4+\mu\theta)(4+3\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$-\frac{2\mu\theta(4+\mu\theta)(4+3\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$-\frac{2\mu\theta(4+\mu\theta)(4+3\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$-\frac{2\mu\theta(4+\mu\theta)(4+3\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$

	xxyy	xyxy	xxyz	yzxx
c_0	$\frac{1}{256}$	$\frac{1}{256}$	$\frac{1}{256}$	$\frac{1}{256}$
c_1	$\frac{3}{256(1+\mu\theta)}$	$-\frac{1}{256(1+\mu\theta)}$	$-\frac{1}{256(1+\mu\theta)}$	$\frac{3}{256(1+\mu\theta)}$
c_2	$-\frac{2}{256(1+\mu\theta)}$	$\frac{2}{256(1+\mu\theta)}$	$-\frac{2}{256(1+\mu\theta)}$	$-\frac{2}{256(1+\mu\theta)}$
c_3	$-\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	$\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$
c_4	$\frac{1}{256(1+\mu\theta)}$	$\frac{1}{256(1+\mu\theta)}$	$\frac{1}{256(1+\mu\theta)}$	$-\frac{3}{256(1+\mu\theta)}$
c_5	$-\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	$\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$
c_6	$\frac{9}{256(1+\mu\theta)^2}$	$\frac{1}{256(1+\mu\theta)^2}$	$-\frac{1}{256(1+\mu\theta)^2}$	$-\frac{3}{256(1+\mu\theta)^2}$
c_7	$-\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$	$\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$
c_8	$-\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$	$\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$	$\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$	$\frac{32}{256(1+\mu\theta)(2+\mu\theta)^2}$
c_9	$\frac{2\mu\theta(4+\mu\theta)^2}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$\frac{\mu\theta(32+40\mu\theta+10\mu^2\theta^2)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$\frac{2\mu^2\theta^2(4+\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$\frac{2\mu^2\theta^2(4+\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$

	xyxz	yxxz	xyzw
c_0	$\frac{1}{256}$	$\frac{1}{256}$	$\frac{1}{256}$
c_1	$-\frac{1}{256(1+\mu\theta)}$	$-\frac{1}{256(1+\mu\theta)}$	$-\frac{1}{256(1+\mu\theta)}$
c_2	$-\frac{2}{256(1+\mu\theta)}$	$\frac{2}{256(1+\mu\theta)}$	$-\frac{2}{256(1+\mu\theta)}$
c_3	$\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	$-\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	$\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$
c_4	$\frac{1}{256(1+\mu\theta)}$	$-\frac{3}{256(1+\mu\theta)}$	$-\frac{3}{256(1+\mu\theta)}$
c_5	$-\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	$\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$	$\frac{4}{256(1+\mu\theta)(2+\mu\theta)}$
c_6	$\frac{1}{256(1+\mu\theta)^2}$	$\frac{1}{256(1+\mu\theta)^2}$	$\frac{1}{256(1+\mu\theta)^2}$
c_7	0	0	$\frac{8}{256(1+\mu\theta)(2+\mu\theta)}$
c_8	0	0	$-\frac{16}{256(1+\mu\theta)(2+\mu\theta)^2}$
c_9	$-\frac{\mu^2\theta^2(4+2\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$-\frac{\mu^2\theta^2(4+2\mu\theta)}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$	$\frac{2\mu^3\theta^3}{256(1+\mu\theta)^2(2+\mu\theta)^2(3+\mu\theta)}$

Appendix B.

Below are the expressions for the variances of the parameters $y_i = x_i^{2\mu}$, $i = 1, 2, 3$, for both the symmetric and the asymmetric species trees. In all of the expressions below, $\sigma_i = var(\hat{p}_i)$ and $\sigma_{i,j} = cov(\hat{p}_i, \hat{p}_j)$, where p_i refers to the relevant entry in $\mathbf{p_s}$ or $\mathbf{p_a}$ for the symmetric or asymmetric species trees, respectively.

B.1. Symmetric Case

$$
\begin{aligned}
Var(\hat{y}_1) = 16(1+\mu\theta)^2(&\sigma_1 + 4\sigma_2 + 9\sigma_5 + 4\sigma_6 + 36\sigma_3 + 4\sigma_4 + 64\sigma_8 + 4\sigma_9 + 36\sigma_7 \\
&- 4\sigma_{1,2} + 6\sigma_{1,5} - 4\sigma_{1,6} + 12\sigma_{1,3} - 4\sigma_{1,4} - 16\sigma_{1,8} - 4\sigma_{1,9} + 12\sigma_{1,7} - 12\sigma_{2,5} + 8\sigma_{2,6} \\
&- 24\sigma_{2,3} + 8\sigma_{2,4} + 32\sigma_{2,8} + 8\sigma_{2,9} - 24\sigma_{2,7} - 12\sigma_{5,6} + 36\sigma_{3,5} - 12\sigma_{4,5} - 48\sigma_{5,8} - 12\sigma_{5,9} \\
&+ 36\sigma_{5,7} - 24\sigma_{3,6} + 8\sigma_{4,6} + 32\sigma_{6,8} + 8\sigma_{6,9} - 24\sigma_{6,7} - 24\sigma_{3,4} - 96\sigma_{3,8} - 24\sigma_{3,9} + 72\sigma_{3,7} \\
&+ 32\sigma_{4,8} + 8\sigma_{4,9} - 24\sigma_{4,7} + 32\sigma_{8,9} - 96\sigma_{7,8} - 24\sigma_{7,9})
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{y}_2) = 16(1+\mu\theta)^2(&\sigma_1 + 36\sigma_2 + 9\sigma_5 + 36\sigma_6 + 4\sigma_3 + 4\sigma_4 + 64\sigma_8 + 4\sigma_9 + 4\sigma_7 \\
&+ 12\sigma_{1,2} + 6\sigma_{1,5} + 12\sigma_{1,6} - 4\sigma_{1,3} - 4\sigma_{1,4} - 16\sigma_{1,8} - 4\sigma_{1,9} - 4\sigma_{1,7} + 36\sigma_{2,5} + 72\sigma_{2,6} \\
&- 24\sigma_{2,3} + 8\sigma_{2,4} + 32\sigma_{2,8} + 8\sigma_{2,9} - 24\sigma_{2,7} + 36\sigma_{5,6} - 12\sigma_{3,5} - 12\sigma_{4,5} - 48\sigma_{5,8} - 12\sigma_{5,9} \\
&- 12\sigma_{5,7} - 24\sigma_{3,6} + 8\sigma_{4,6} + 32\sigma_{6,8} + 8\sigma_{6,9} - 24\sigma_{6,7} - 24\sigma_{3,4} - 96\sigma_{3,8} - 24\sigma_{3,9} + 8\sigma_{3,7} \\
&+ 32\sigma_{4,8} + 8\sigma_{4,9} + 8\sigma_{4,7} + 32\sigma_{8,9} + 32\sigma_{7,8} + 8\sigma_{7,9})
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{y}_3) = 16(1+\mu\theta)^2(&\sigma_1 + 4\sigma_2 + \sigma_5 + 4\sigma_6 + 4\sigma_3 + 4\sigma_4 + 4\sigma_9 + 4\sigma_7 \\
&+ 4\sigma_{1,2} - 2\sigma_{1,5} - 4\sigma_{1,6} + 4\sigma_{1,3} + 4\sigma_{1,4} - 4\sigma_{1,9} - 4\sigma_{1,7} - 4\sigma_{2,5} - 8\sigma_{2,6} \\
&+ 8\sigma_{2,3} + 8\sigma_{2,4} - 8\sigma_{2,9} - 8\sigma_{2,7} + 4\sigma_{5,6} - 4\sigma_{3,5} - 4\sigma_{4,5} + 4\sigma_{5,9} \\
&+ 4\sigma_{5,7} - 8\sigma_{3,6} - 8\sigma_{4,6} + 8\sigma_{6,9} + 8\sigma_{6,7} + 8\sigma_{3,4} - 8\sigma_{3,9} - 8\sigma_{3,7} \\
&- 8\sigma_{4,9} - 8\sigma_{4,7} + 8\sigma_{7,9})
\end{aligned}
$$

B.2. Asymmetric Case

$$
\begin{aligned}
Var(\hat{y}_1) = (16/9)(1+\mu\theta)^2(&9\sigma_1 + 49\sigma_2 + 64\sigma_6 + 64\sigma_7 + 100\sigma_3 + 25\sigma_5 + 100\sigma_9 + 36\sigma_{11} + 81\sigma_4 + 144\sigma_{10} + 324\sigma_8 \\
&- 42\sigma_{1,2} + 48\sigma_{1,6} - 48\sigma_{1,7} + 60\sigma_{1,3} - 30\sigma_{1,5} - 60\sigma_{1,9} - 36\sigma_{1,11} + 54\sigma_{1,4} - 72\sigma_{1,10} + 108\sigma_{1,8} \\
&- 112\sigma_{2,6} + 112\sigma_{2,7} - 140\sigma_{2,3} + 70\sigma_{2,5} + 140\sigma_{2,9} + 84\sigma_{2,11} - 126\sigma_{2,4} + 168\sigma_{2,10} - 252\sigma_{2,8} - 128\sigma_{6,7} \\
&+ 160\sigma_{3,6} - 80\sigma_{5,6} - 160\sigma_{6,9} - 96\sigma_{6,11} + 144\sigma_{4,6} - 192\sigma_{6,10} + 288\sigma_{6,8} - 160\sigma_{3,7} + 80\sigma_{5,7} \\
&+ 160\sigma_{7,9} + 96\sigma_{7,11} - 144\sigma_{4,7} + 192\sigma_{7,10} - 288\sigma_{7,8} - 100\sigma_{3,5} - 200\sigma_{3,9} - 120\sigma_{3,11} + 180\sigma_{3,4} \\
&- 240\sigma_{3,10} + 360\sigma_{3,8} + 100\sigma_{5,9} + 60\sigma_{5,11} - 90\sigma_{4,5} + 120\sigma_{5,10} - 180\sigma_{5,8} + 120\sigma_{9,11} - 180\sigma_{4,9} \\
&+ 240\sigma_{9,10} - 360\sigma_{8,9} - 108\sigma_{4,11} + 144\sigma_{10,11} - 216\sigma_{8,11} - 216\sigma_{4,10} + 324\sigma_{4,8} - 432\sigma_{8,10})
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{y}_2) = (16/9)(1+\mu\theta)^2(&9\sigma_1 + 25\sigma_2 + 16\sigma_6 + 64\sigma_7 + 4\sigma_3 + 49\sigma_5 + 100\sigma_9 + 36\sigma_{11} + 81\sigma_4 + 144\sigma_{10} + 36\sigma_8 \\
&+ 30\sigma_{1,2} - 24\sigma_{1,6} - 48\sigma_{1,7} - 12\sigma_{1,3} + 42\sigma_{1,5} - 60\sigma_{1,9} - 36\sigma_{1,11} + 54\sigma_{1,4} + 72\sigma_{1,10} - 36\sigma_{1,8} \\
&- 40\sigma_{2,6} - 80\sigma_{2,7} - 20\sigma_{2,3} + 70\sigma_{2,5} - 100\sigma_{2,9} - 60\sigma_{2,11} + 90\sigma_{2,4} + 120\sigma_{2,10} - 60\sigma_{2,8} + 64\sigma_{6,7} \\
&+ 16\sigma_{3,6} - 56\sigma_{5,6} + 80\sigma_{6,9} + 48\sigma_{6,11} - 72\sigma_{4,6} - 96\sigma_{6,10} + 48\sigma_{6,8} + 32\sigma_{3,7} - 112\sigma_{5,7} \\
&+ 160\sigma_{7,9} + 96\sigma_{7,11} - 144\sigma_{4,7} - 192\sigma_{7,10} + 96\sigma_{7,8} - 28\sigma_{3,5} + 40\sigma_{3,9} + 24\sigma_{3,11} - 36\sigma_{3,4} \\
&- 48\sigma_{3,10} + 24\sigma_{3,8} - 140\sigma_{5,9} - 84\sigma_{5,11} + 126\sigma_{4,5} + 168\sigma_{5,10} - 84\sigma_{5,8} + 120\sigma_{9,11} - 180\sigma_{4,9} \\
&- 240\sigma_{9,10} + 120\sigma_{8,9} - 108\sigma_{4,11} - 144\sigma_{10,11} + 72\sigma_{8,11} + 216\sigma_{4,10} - 108\sigma_{4,8} - 144\sigma_{8,10})
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{y}_3) = (16/9)(1+\mu\theta)^2(&9\sigma_1 + 100\sigma_2 + var6 + 4\sigma_7 + 25\sigma_3 + 4\sigma_5 + 16\sigma_9 + 36\sigma_{11} + 9\sigma_4 + 144\sigma_{10} + 36\sigma_8 \\
&+ 60\sigma_{1,2} + 6\sigma_{1,6} + 12\sigma_{1,7} + 30\sigma_{1,3} + 12\sigma_{1,5} + 24\sigma_{1,9} - 36\sigma_{1,11} - 18\sigma_{1,4} - 72\sigma_{1,10} - 36\sigma_{1,8} \\
&+ 20\sigma_{2,6} + 40\sigma_{2,7} + 100\sigma_{2,3} + 40\sigma_{2,5} + 80\sigma_{2,9} - 120\sigma_{2,11} - 60\sigma_{2,4} - 240\sigma_{2,10} - 120\sigma_{2,8} + 4\sigma_{6,7} \\
&+ 10\sigma_{3,6} + 4\sigma_{5,6} + 8\sigma_{6,9} - 12\sigma_{6,11} - 6\sigma_{4,6} - 24\sigma_{6,10} - 12\sigma_{6,8} + 20\sigma_{3,7} + 8\sigma_{5,7} \\
&+ 16\sigma_{7,9} - 24\sigma_{7,11} - 12\sigma_{4,7} - 48\sigma_{7,10} - 24\sigma_{7,8} + 20\sigma_{3,5} + 40\sigma_{3,9} - 60\sigma_{3,11} - 30\sigma_{3,4} \\
&- 120\sigma_{3,10} - 60\sigma_{3,8} + 16\sigma_{5,9} - 24\sigma_{5,11} - 12\sigma_{4,5} - 48\sigma_{5,10} - 24\sigma_{5,8} - 48\sigma_{9,11} - 24\sigma_{4,9} \\
&- 96\sigma_{9,10} - 48\sigma_{8,9} + 36\sigma_{4,11} + 144\sigma_{10,11} + 72\sigma_{8,11} + 72\sigma_{4,10} + 36\sigma_{4,8} + 144\sigma_{8,10})
\end{aligned}
$$

Appendix C. Results for Three Taxa

In this section, we provide results analogous to those presented above for the case of three taxa. Our formulation draws on the work of Zhu and Yang (2021), who provided expressions for the site pattern probabilities of a three-taxon species tree under the JC69 model [10] in terms of the branch lengths and effective population size parameters. In their formulation, each branch was assigned a separate effective population size parameter. As we did in the case of four taxa, we assume that the effective population size is constant throughout the tree, and thus there is only a single effective population size parameter, θ.

We assume that the three-taxon species tree is of the form (a:τ_1, b:τ_1):$\tau_2 - \tau_1$,c:τ_2); . We note that this notation matches what we used in the four-taxon case presented above, but differs from that used by Zhu and Yang (2021). There are 4 unique site pattern probabilities for three taxa, which we denote

$$p_{xxx}, \quad p_{xxy}, \quad p_{yxx} = p_{xyx}, \quad p_{xyz}$$

We can then solve Equation (18) in Zhu and Yang (2021) to get the following expressions:

$$\begin{aligned}
\widehat{e^{-\frac{8}{3}\tau_1}} &= \left(\frac{4}{3} + \frac{16}{9}\theta\right)\hat{p}_{xxx} + \left(\frac{4}{3} + \frac{16}{9}\theta\right)\hat{p}_{xxy} - \left(\frac{1}{3} + \frac{4}{9}\theta\right) \\
\widehat{e^{-\frac{8}{3}\tau_2}} &= \left(\frac{4}{3} + \frac{16}{9}\theta\right)\hat{p}_{xxx} + \left(\frac{2}{3} + \frac{8}{9}\theta\right)\hat{p}_{yxx} + \left(\frac{2}{3} + \frac{8}{9}\theta\right)\hat{p}_{xyx} - \left(\frac{1}{3} + \frac{4}{9}\theta\right) \\
\hat{\tau}_1 &= -\frac{3}{8}\ln\left[\left(\frac{4}{3} + \frac{16}{9}\theta\right)\hat{p}_{xxx} + \left(\frac{4}{3} + \frac{16}{9}\theta\right)\hat{p}_{xxy} - \left(\frac{1}{3} + \frac{4}{9}\theta\right)\right] \\
\hat{\tau}_2 &= -\frac{3}{8}\ln\left[\left(\frac{4}{3} + \frac{16}{9}\theta\right)\hat{p}_{xxx} + \left(\frac{2}{3} + \frac{8}{9}\theta\right)\hat{p}_{yxx} + \left(\frac{2}{3} + \frac{8}{9}\theta\right)\hat{p}_{xyx} - \left(\frac{1}{3} + \frac{4}{9}\theta\right)\right]
\end{aligned}$$

As in the case of four taxa, variances for the first two expressions can be derived and used to construct confidence intervals for τ_1 and τ_2. Following the same process as in the four-taxon case, we can derive the following expressions for the variances:

$$
\begin{aligned}
\widehat{Var(e^{-\frac{8}{3}\tau_1})} &= \left(\frac{4}{3}+\frac{16}{9}\theta\right)^2 Var(\hat{p}_{xxx}) + \left(\frac{4}{3}+\frac{16}{9}\theta\right)^2 Var(\hat{p}_{xxy}) + 2\left(\frac{4}{3}+\frac{16}{9}\theta\right)^2 Cov(\hat{p}_{xxx},\hat{p}_{xxy}) \\
\widehat{Var(e^{-\frac{8}{3}\tau_2})} &= \left(\frac{4}{3}+\frac{16}{9}\theta\right)^2 Var(\hat{p}_{xxx}) + \left(\frac{2}{3}+\frac{8}{9}\theta\right)^2 Var(\hat{p}_{yxx}) + \left(\frac{2}{3}+\frac{8}{9}\theta\right)^2 Var(\hat{p}_{xyx}) \\
&+ 2\left(\frac{4}{3}+\frac{16}{9}\theta\right)\left(\frac{2}{3}+\frac{8}{9}\theta\right) Cov(\hat{p}_{xxx},\hat{p}_{yxx}) + 2\left(\frac{4}{3}+\frac{16}{9}\theta\right)\left(\frac{2}{3}+\frac{8}{9}\theta\right) Cov(\hat{p}_{xxx},\hat{p}_{xyx}) \\
&+ 2\left(\frac{2}{3}+\frac{8}{9}\theta\right)^2 Cov(\hat{p}_{yxx},\hat{p}_{xyx})
\end{aligned}
$$

With the estimators of τ_1, τ_2, and their variances in hand, we can repeat the simulation studies undertaken in the four-taxon case. Specifically, we run simulations corresponding to Settings 1-4 in Section 4. We also considered the application to species delimitation described in Section 5. All results are presented in the Supplemental Material, and we include an R Markdown file with all of the code in the material available for download from GitHub.