



Identifiability and Reconstructibility of Species Phylogenies Under a Modified Coalescent

Colby Long¹ · Laura Kubatko²

Received: 21 April 2017 / Accepted: 11 June 2018 / Published online: 20 June 2018
© Society for Mathematical Biology 2018

Abstract

Coalescent models of evolution account for incomplete lineage sorting by specifying a species tree parameter which determines a distribution on gene trees, and consequently, a site pattern probability distribution. It has been shown that the unrooted topology of the species tree parameter of the multispecies coalescent is generically identifiable, and a reconstruction method called SVDQuartets has been developed to infer this topology. In this paper, we describe a modified multispecies coalescent model that allows for varying effective population size and violations of the molecular clock. We show that the unrooted topology of the species tree parameter for these models is generically identifiable and that SVDQuartets can still be used to infer this topology.

Keywords Molecular clock · SVDQuartets · Multispecies Coalescent

1 Introduction

The goal of phylogenetics is to reconstruct the evolutionary history of a group of species from biological data. Most often, the data available are the aligned DNA sequences of the species under consideration. The descent of these species from a common ancestor is represented by a rooted phylogenetic tree which we call the *species tree*. However, it is well known that due to various biological phenomena, such as horizontal gene transfer and incomplete lineage sorting, the ancestry of individual genes will not necessarily match the tree of the species in which they reside (Pamilo and Nei 1988; Syvanen 1994; Maddison 1997). There are various phylogenetic reconstruction meth-

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11538-018-0456-9>) contains supplementary material, which is available to authorized users.

✉ Colby Long
long.1579@mbi.osu.edu

¹ Mathematical Biosciences Institute, Columbus, USA

² Departments of Statistics and Evolution, Ecology, and Organismal Biology,
The Ohio State University, Columbus, USA

ods that account for this discrepancy in different ways. One approach is to reconstruct individual *gene trees* by some method and then utilize this information to infer the original species tree (Liu et al. 2009, 2010; Wu 2012; Mirarab et al. 2014; Mirarab and Warnow 2015).

The multispecies coalescent model incorporates incomplete lineage sorting directly. The tree parameter of the model is the species tree, an n -leaf rooted equidistant tree with branch lengths. The species tree yields a distribution on possible gene trees along which evolution is modeled by a κ -state substitution model. For a fixed choice of parameters, the multispecies coalescent returns a probability distribution on the κ^n possible n -tuples of states that may be observed. In order to infer the species tree from data, one searches for model parameters yielding a distribution close to that observed, using, for example, maximum likelihood.

In Chifman and Kubatko (2015), the authors show that given a probability distribution from the multispecies coalescent model, it is possible to infer the unrooted topology of the species tree parameter. Unrooting the species tree and restricting to any four-element subset of the leaves yield an unrooted four-leaf binary phylogenetic tree called a *quartet*. For a given label set, there are only three possible quartets which each induce a flattening of the probability tensor. Given a probability distribution arising from the multispecies coalescent, the flattening matrix corresponding to the quartet compatible with the species tree will be rank $\binom{\kappa+1}{2}$ or less while the other two will generically have rank strictly greater than this value. Since the topology of an unrooted tree is uniquely determined by quartets Semple and Steel (2003), these flattening matrices can be used to determine the unrooted topology of the species tree exactly. Of course, empirical and even simulated data produced by the multispecies coalescent will only approximate the distribution arising from the model. Therefore, the same authors also proposed a method called SVDQuartets Chifman and Kubatko (2014), which uses singular value decomposition to infer each quartet topology by determining which of the flattening matrices is closest to the set of rank $\binom{\kappa+1}{2}$ matrices.

The method of SVDQuartets offers several advantages over other existing phylogenetic reconstruction methods. For example, it accounts for incomplete lineage sorting and is computationally much less expensive than Bayesian methods achieving the same level of accuracy. It is often underappreciated that this reconstruction method can be used to recover the species tree for several different underlying nucleotide substitution models without any modifications. It was shown in Chifman and Kubatko (2015) that the method of SVDQuartets is applicable when the underlying model for the evolution of sequence data along the gene trees is the four-state general time-reversible (GTR) model or any of the commonly used submodels thereof (e.g., JC69, K2P, K3P, F81, HKY85, TN93). Thus, the method does not require any a priori assumptions about the underlying nucleotide substitution process other than time reversibility.

In this paper, we show that the method of SVDQuartets has more theoretical robustness even than has already been shown. We will specifically focus on the case where the underlying nucleotide substitution model is one of the four-state models most widely used in phylogenetics. We describe several modifications to the classical multispecies coalescent model to allow for more realistic mechanisms of evolution. For example, we remove the assumption of a molecular clock by removing the restriction that the species tree be equidistant. We also allow the effective population size

to vary on each branch of the species tree. Remarkably, we show that the unrooted topology of the species tree parameter of these modified models is still identifiable and that SVDQuartets is still an appropriate reconstruction method. Thus, despite the introduction of several parameters, effective and efficient methods for reconstructing the unrooted topology of the species tree for these modified coalescent models are already available off the shelf and implemented in PAUP* (Swofford 2002).

In Sect. 2, we review the classical multispecies coalescent model and discuss some of its limitations in modeling certain biological phenomena. We then describe several modifications to the classical model to remedy these weaknesses. In Sect. 3, we establish the theoretical properties of identifiability for these families of modified coalescent models. Finally, in Sect. 4, we describe why SVDQuartets is a strong candidate for reconstructing the species tree under the multispecies coalescent and propose several other modifications that could be made to the multispecies coalescent.

2 The Multispecies Coalescent

2.1 Coalescent Models of Evolution

In this section, we briefly review the multispecies coalescent model and explain how the model yields a probability distribution on nucleotide site patterns. As our main results will parallel those found in Chifman and Kubatko (2015), we will import much of the notation from that paper and refer the reader there for a more thorough description of the model.

The Wright-Fisher model from population genetics models the convergence of multiple lineages backward in time toward a common ancestor. Beginning with j lineages from the current generation, the model assumes discrete generations with constant effective population size N . In each generation, each lineage is assigned a parent uniformly from the previous generation. For diploid species, there are $2N$ copies of each gene in each generation, and thus the probability of selecting any particular gene as a parent is $\frac{1}{2N}$. Two lineages are said to coalesce when they share the same parent in a particular generation.

As an example, if we begin with two lineages in the same species, the probability they have the same parent in the previous generation, and hence coalesce, is $\frac{1}{2N}$ and the probability that they do not coalesce in this generation is $(1 - \frac{1}{2N})$. Therefore, the probability that two lineages coalesce in exactly the i th previous generation is given by

$$\left(\frac{1}{2N}\right)\left(1 - \frac{1}{2N}\right)^{i-1}.$$

For large N , the time at which the two lineages coalesce, t , approximately follows an exponential distribution with rate $(2N)^{-1}$, where time is measured in number of generations. Every $2N$ generation is called a *coalescent unit*, and time is typically measured in these units to simplify the formulas for time to coalescence. However,

in this paper, we will introduce separate effective population size parameters for each branch of the species trees. So that our timescale is consistent across the tree we will work in generations rather than coalescent units. In these units, for j lineages, the time to the next coalescent event t has probability density,

$$f(t) = \frac{j(j-1)}{2} \left(\frac{1}{2N}\right) \exp\left(-\frac{j(j-1)}{2} \left(\frac{1}{2N}\right) t\right), \quad t > 0. \quad (1)$$

This is typically referred to as *Kingman's coalescent* (Kingman 1982a,b,c; Tajima 1983; Tavaré 1984; Takahata and Nei 1985).

The multispecies coalescent is based on the same framework, but we assume that the species tree of the sampled taxa is known. We let S denote the topology (without branch lengths) of the n -leaf rooted binary phylogenetic species tree. The tips of S represent distinct species and are labeled by uppercase letters. We assume here that one lineage is sampled from each species, and we label each lineage by the lowercase letter corresponding to the species from which it is sampled. We use e_X to denote the branch of S that is ancestral to exactly the species in X . The vector τ specifies branch lengths where τ_X is the length of e_X . Thus, (S, τ) denotes a rooted species tree with branch lengths. In the classical multispecies coalescent, the entries of τ are chosen so that (S, τ) is *equidistant*, meaning that the length of the path from the root to any tip of the species tree is the same. This is commonly referred to as the molecular clock assumption, and in what follows, we refer to the classical multispecies coalescent as the *equidistant coalescent*. For example, for the four-leaf species tree depicted in Fig. 1a, $\tau = (\tau_A, \tau_B, \tau_C, \tau_D, \tau_{AB}, \tau_{CD})$ and the entries satisfy $\tau_A = \tau_B$, $\tau_C = \tau_D$, and $\tau_A + \tau_{AB} = \tau_C + \tau_{CD}$. Later, we will introduce different effective population sizes in each population and we will use N_X to denote the size of the population in e_X .

Once this species tree is fixed, the multispecies coalescent gives a probability density on possible gene trees, where here we use the term gene tree to mean both the topology and the branch lengths. All of the same assumptions above apply, except that it is now impossible for two lineages to coalesce if they are not part of the same population. Hence, lineages may only coalesce if they are in the same branch of S . We use the concept of a *coalescent history*, h (see, e.g., Degnan and Salter 2005) to indicate a particular sequence of coalescent events as well as the populations in which they occur (but not the precise times of the events). There are only finitely many possible coalescent histories compatible with S , and we call the set of all such histories \mathcal{H} . We denote the topology of a rooted n -leaf binary phylogenetic gene tree by G and let $\mathbf{t} = (t_1, \dots, t_{n-1})$ be the vector that encodes the coalescent times. Thus, in the context of a specific species tree and history, (G, \mathbf{t}) encodes a gene tree with branch lengths. As for the species tree, we measure all branch lengths in units of generations. Note that any given history corresponds to infinitely many gene trees, though all will have the same topology. Likewise, a particular gene tree topology may correspond to only one history or (finitely) many histories. Figure 1 gives an example of a single gene tree topology with two distinct histories. Note, however, that in Fig. 1a, b, there are infinitely many choices for the values of t_1 , t_2 , and t_3 that satisfy the constraints of each history.

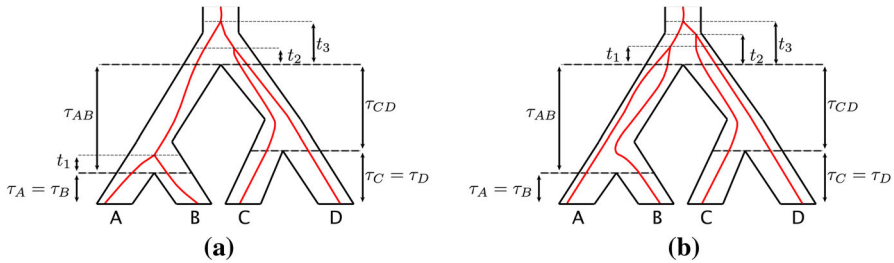


Fig. 1 A species tree and two different coalescent histories that result in the same gene tree topology. The coalescent times t_j are measured from the most recent speciation event (looking backwards in time). **a** A gene tree with topology G and history h . **b** A gene tree with topology G and history h'

For a particular history and a particular species tree (S, τ) , we can compute the probability density for gene trees (G, \mathbf{t}) with that history explicitly under the multi-species coalescent model. We denote this gene tree density by $f_h((G, \mathbf{t})|(S, \tau))$. We demonstrate below how this is done for the history in Fig. 1a.

Example 2.1 Let (S, τ) be the four-leaf species tree depicted in Fig. 1a, and let h refer to the coalescent history in which the following events occur in order (looking backward from the present):

- (1) Lineages a and b coalesce in the population ancestral to A and B .
- (2) Lineages c and d coalesce in the population above the root.
- (3) Lineages ab and cd coalesce in the population above the root.

The probability of observing a gene tree with history h for the species tree (S, τ) under the multispecies coalescent model can be found by integrating over all possible times at which the coalescent events consistent with h may occur. Note that the integrals must be taken with respect to the boundaries for the coalescent events specified by the history. Therefore, each history will have a unique region of integration, and each must be considered separately. For history h shown in Fig. 1a, we have

$$\int_{\mathbf{t}} f_h((G, \mathbf{t})|(S, \tau)) d\mathbf{t} = \int_0^\infty \int_0^{t_3} \int_0^{\tau_{AB}} \exp\left(\frac{-\tau_{CD}}{2N}\right) \left(\left(\frac{1}{2N}\right) \exp\left(\frac{-t_1}{2N}\right)\right) \left(\left(\frac{1}{2N}\right) \exp\left(\frac{-3t_2}{2N}\right)\right) \left(\left(\frac{1}{2N}\right) \exp\left(\frac{-(t_3 - t_2)}{2N}\right)\right) dt_1 dt_2 dt_3.$$

We assume that the distribution of times to coalescent events is given by Kingman’s coalescent as in Eq. (1). We compute $f_h((G, \mathbf{t})|(S, \tau))$ under the assumption that evolution occurs independently in each branch of the species tree. Thus, we multiply the contributions to the density of the events occurring in each species tree branch to obtain the probability density of the history. For example, the first term of $f_h((G, \mathbf{t})|(S, \tau))$ is equal to

$$1 - \int_0^{\tau_{CD}} \left(\frac{1}{2N}\right) \exp\left(\frac{-t}{2N}\right) dt,$$

the probability that lineages c and d do not coalesce in the population ancestral to C and D . The second term is the probability density for the event that lineages a and b coalesce at time t_1 . The third term is the probability density for the event that lineages c and d coalesce at time t_2 . Finally, the last term is the probability density for the event that the newly formed lineages ab and cd coalesce at time $t_3 - t_2$ (measured relative to the time when lineages c and d coalesced). Notice that the coalescent history h' depicted in Fig. 1b results in the same gene tree, but the density function $f_{h'}((G, \mathbf{t})|(S, \boldsymbol{\tau}))$ will not be the same as $f_h((G, \mathbf{t})|(S, \boldsymbol{\tau}))$.

For a fixed gene tree at a given locus, we model the evolution along this gene tree as a continuous-time homogenous Markov process according to a nucleotide substitution model. The model gives a probability distribution on the set of all 4^n possible n -tuples of observed states at the leaves of (G, \mathbf{t}) . We can write the probability of observing the state (i_1, \dots, i_n) as $p_{i_1 \dots i_n}^*|(G, \mathbf{t})$. Precisely how this distribution is calculated is described in Chifman and Kubatko (2015). Here, we sketch the relevant details needed to introduce the modified multispecies coalescent model described in the next section.

For a four-state substitution model, there is a 4×4 instantaneous rate matrix Q where the entry Q_{ij} encodes the rate of conversion from state i to state j . To compute the probability of observing a particular state at the leaves, we associate with each vertex v a random variable X_v with state space equal to the set of four possible states. The distribution of states at the root vertex is $\boldsymbol{\pi} = (\pi_A, \pi_G, \pi_C, \pi_T)$ where $\boldsymbol{\pi}$ is the stationary distribution of the rate matrix Q . Letting t_e be the length of edge $e = uv$, $P(t_e) = e^{Qt_e}$ is the matrix of transition probabilities along that edge. That is, $P_{ij}(t_e) = P(X_v = j | X_u = i)$. Given an assignment of states to each vertex of the tree, we can compute the probability of observing this state using the Markov property and the appropriate entries of the transition matrices. To determine the probability of observing a particular state at the leaves, we marginalize over all possible states of the internal nodes.

In this paper, we are primarily interested in four-state models of DNA evolution where the four states correspond to the DNA bases. Different phylogenetic models place different restrictions on the entries of the rate matrices. The results that we prove in the next section will apply when the underlying nucleotide substitution model is any of the commonly used four-state time-reversible models. As an example, the rate matrices for two of these models, the Kimura three-parameter model (K3P) and the four-state general time-reversible model (GTR), are given in Fig. 2. We note here that because these models are time reversible, the location of the root in each gene tree is unidentifiable from the site pattern probability distribution for that gene tree Felsenstein (1981). In subsequent sections, we will introduce and describe similar results for the JC+I+ Γ model that allows for invariable sites and gamma-distributed rates across sites.

Now, given a species tree $(S, \boldsymbol{\tau})$ and a choice of nucleotide substitution model, let $p_{i_1 \dots i_n}^*|(S, \boldsymbol{\tau})$ be the probability of observing the site pattern $i_1 \dots i_n$ at the tips of $(S, \boldsymbol{\tau})$. To compute $p_{i_1 \dots i_n}^*|(S, \boldsymbol{\tau})$, we must consider the contribution of each history to the site pattern probability distribution by integrating over branch lengths. So that we may write the formulas explicitly, and we first consider the contribution of gene trees matching a particular coalescent history,

$$\begin{matrix}
 \begin{pmatrix} * & a & b & c \\ a & * & c & b \\ b & c & * & a \\ c & b & a & * \end{pmatrix} &
 \begin{pmatrix} * & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & * & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & * & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & * \end{pmatrix} \\
 \text{(a)} & \text{(b)}
 \end{matrix}$$

Fig. 2 Rate matrices for two commonly used models in phylogenetics. The diagonal entries are chosen so that the row sums are equal to zero. In the K3P model, the root distribution is uniform. **a** Kimura three-parameter model (K3P). **b** Four-state general time-reversible model (GTR)

$$p_{i_1 \dots i_n | h, (S, \tau)} = \int_{\mathbf{t}} p_{i_1 \dots i_n | (G, \mathbf{t})}^* f_h((G, \mathbf{t}) | (S, \tau)) \, d\mathbf{t}.$$

As noted previously, there will be finitely many histories for any given species tree (S, τ) , and summing over these gives the probability of observing the site pattern $i_1 \dots i_n$ at the tips of the species tree (S, τ) ,

$$\begin{aligned}
 p_{i_1 \dots i_n | (S, \tau)} &= \sum_{h \in \mathcal{H}} \int_{\mathbf{t}} p_{i_1 \dots i_n | (G, \mathbf{t})}^* f_h((G, \mathbf{t}) | (S, \tau)) \, d\mathbf{t} \\
 &= \sum_{h \in \mathcal{H}} p_{i_1 \dots i_n | h, (S, \tau)}.
 \end{aligned}$$

Note again that the bounds of integration in each term of the sum will depend on the history being considered.

2.2 A Modified Coalescent

In this section, we introduce various ways that we might alter the multispecies coalescent to better reflect the evolutionary process. Recall that the length of the path from the root of the species tree to each tip is the total number of generations that have occurred between the species at the root and that at the tip. Since the length of a generation may vary for different species Martin and Palumbi (1993), it may be desirable to allow the lengths of the paths from the root to each tip to differ. Therefore, we first consider expanding the allowable set of branch lengths so that (S, τ) is not required to be equidistant.

Fix a nucleotide substitution model. Let $\mathcal{C}(S) \subseteq \Delta^{4^n - 1}$ be the set of site pattern probability distributions obtained from the equidistant multispecies coalescent model on the n -leaf topological rooted tree S . Let $\mathcal{C}_n \subseteq \Delta^{4^n - 1}$ denote the set of all distributions obtained by allowing (S, τ) to be any equidistant n -leaf rooted tree. If (S, τ) is not required to be equidistant, this removes the assumption of a molecular clock and we refer to this model as the *clockless coalescent*. The set of site pattern probabilities obtained from a single species tree topology in the clockless coalescent is $\mathcal{C}^*(S)$, and the set of distributions obtained by allowing (S, τ) to be any n -leaf rooted tree (not necessarily equidistant) is \mathcal{C}_n^* .

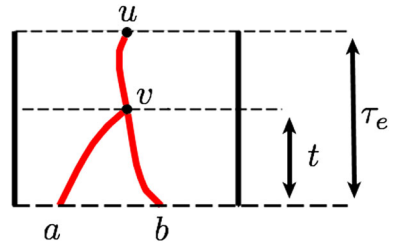
We can also account for the fact that the effective population size, N , may vary for different species Charlesworth (2009) by introducing a separate effective population size parameter for each internal branch of the species tree. We call this model the p -coalescent and denote the set of all site pattern probabilities arising from the model as $\mathcal{C}(S, N)$. Note that here we consider the species tree (S, τ) to be equidistant. In analogy to our notation from above, we use $\mathcal{C}_n(N)$ to denote the set of all site pattern probability distributions obtained from the p -coalescent and use $\mathcal{C}^*(S, N)$ and $\mathcal{C}_n^*(N)$ for the clockless p -coalescent.

Since we assume that coalescent events do not occur within terminal edges of the species tree, changing the effective population size on the terminal edges does not change the probability distribution on gene trees or the site pattern probabilities arising from a given gene tree. In the next section, we will show that, remarkably, the unrooted topologies of the species tree parameter of the clockless coalescent, p -coalescent, and the clockless p -coalescent are all generically identifiable. Conveniently, from the perspective of reconstruction, we also show that the method of SVDQuartets Chifman and Kubatko (2014) can be used to reconstruct the unrooted topology of the species tree based on a sample from the site pattern probability distribution given by the model.

It is well known that when considering the gene tree distribution from the coalescent model on a rooted tree S , the branch lengths and population sizes are confounded. For example, if a particular branch length is doubled and the population size on that branch halved, this will not affect the gene tree distribution. However, we note that the site pattern probability distributions induced by the clockless coalescent and by the p -coalescent on S are not necessarily equal (i.e., $\mathcal{C}^*(S)$ does not necessarily equal $\mathcal{C}(S, N)$). Some intuition for why these are not necessarily the same can be obtained by comparing each of these modified coalescent models to the equidistant model. We can construct a species tree from the clockless coalescent by beginning with an equidistant species tree and either stretching or contracting certain branches. This alters the gene tree distribution by allowing more or less time, respectively, for coalescent events to occur along the affected branch. However, the probability density of the time to a fixed coalescent event will not necessarily be affected by this change. In contrast, the p -coalescent induces a change in the *rate* of coalescence, as can be seen by examining Eq. (1), which will alter the probability density of the time of a coalescent event that resides in any affected species tree branch. For example, if the branch e in Fig. 3 is from an equidistant tree (S, τ) , increasing τ_e will not affect the probability density of the time to coalescence of lineages a and b , denoted by t in the figure and given by Eq. (1). However, changing the effective population size along this branch will affect the probability density of t , since the effective population size N appears in Eq. (1).

One might be interested in a generalization of the multispecies coalescent model in which a mutation rate is associated with each branch of the species tree. This is biologically realistic in that it would allow for mutation to accumulate at different rates along different branches of the species tree, in response to factors such as variations in climate or other ecological conditions. One might think of modeling this by generalizing the definition of the instantaneous rate matrix Q defined in Section 2.1, so that, rather than associating a single matrix Q with the entire species tree, the lineages within each species tree branch e evolve according to a species tree branch-specific matrix $\rho_e Q$ (ρ_e is a scalar that modifies the mutation rate on branch e). The example

Fig. 3 Two lineages coalescing in a branch of a species tree



below shows that we can obtain the same site pattern probability distribution obtained by scaling Q by ρ_e by instead scaling the length of e and the effective population size in e by ρ_e . This illustrates that a model with a different mutation rate on each species tree edge is subsumed by the clockless p -coalescent.

Example 2.2 Let a and b be two lineages entering a branch e of a species tree as in Fig. 3. Let τ_e be the length of this branch and N be the effective population size parameter. The probability that a and b do not coalesce in e is

$$1 - \int_0^{\tau_e} \left(\frac{1}{2N}\right) \exp\left(\frac{-t}{2N}\right) dt = \exp\left(\frac{-\tau_e}{2N}\right). \tag{2}$$

If a and b coalesce, then we can compute the probability of observing the state xy at a and b under a homogenous Markov model where the rate matrix on the branch e is scaled by a factor ρ_e . We assume the distribution of states at the vertex u is the vector π . Thus, we have,

$$p_{xy} = \sum_{z_1, z_2} \int_0^{\tau_e} \left(\frac{1}{2N}\right) \exp\left(\frac{-t}{2N}\right) \pi_{z_1} \exp(\rho_e Q(\tau_e - t))_{z_1, z_2} \exp(\rho_e Q(t))_{z_2, x} \exp(\rho_e Q(t))_{z_2, y} dt. \tag{3}$$

Instead of scaling the rate matrix Q by ρ_e , we could scale the length of e and the effective population size by ρ_e . Then, the probability that lineages a and b do not coalesce remains unchanged since

$$\exp\left(\frac{-\rho_e \tau_e}{2\rho_e N}\right) = \exp\left(\frac{-\tau_e}{2N}\right).$$

Likewise, the probability of observing state xy is given by the following formula, where we make the substitution $t = \rho_e T$,

$$p_{xy} = \sum_{z_1, z_2} \int_0^{\rho_e \tau_e} \left(\frac{1}{2\rho_e N}\right) \exp\left(\frac{-t}{2\rho_e N}\right) \pi_{z_1} \exp(Q(\rho_e \tau_e - t))_{z_1, z_2} \exp(Q(t))_{z_2, x} \exp(Q(t))_{z_2, y} dt$$

$$= \sum_{z_1, z_2} \int_0^{\tau_e} \left(\frac{1}{2N} \right) \exp\left(\frac{-T}{2N} \right) \pi_{z_1} \exp(\rho_e Q(\tau_e - T))_{z_1, z_2} \exp(\rho_e Q(T))_{z_2, x} \exp(\rho_e Q(T))_{z_2, y} dT.$$

This expression is equal to (3), and thus we have the same distribution of site patterns at the leaves of the tree. Generalizing this example, we can obtain the site pattern probability distribution for a species tree with any branch-specific scaled rate matrices that we desire by appropriately adjusting population sizes and branch lengths across the tree. Thus, we consider only the clockless coalescent, p -coalescent, and clockless p -coalescent models in what follows.

3 Identifiability of the Modified Coalescent

One of the most fundamental concepts in model-based reconstruction is that of identifiability. A model parameter is identifiable if any probability distribution arising from the model uniquely determines the value of that parameter. For the purposes of phylogenetic reconstruction, it is particularly important that the tree parameter of the model be identifiable in order to make consistent inference.

In the following paragraphs, we will use the notation \mathcal{C}_n for the set of site pattern probability distributions obtained by varying the n -leaf tree parameter in the equidistant coalescent model, though the discussion applies equally to \mathcal{C}_n^* , $\mathcal{C}_n(N)$, and $\mathcal{C}_n^*(N)$. To uniquely recover the unrooted topology of the species tree parameter of the n -leaf multispecies coalescent model, we would require that for all n -leaf rooted trees S_1 and S_2 that are topologically distinct when the root vertex of each is suppressed, $\mathcal{C}(S_1) \cap \mathcal{C}(S_2) = \emptyset$. This notion of identifiability is unobtainable in most instances and much stronger than is required in practice. Instead, we often wish to establish *generic identifiability*. A model parameter is generically identifiable if the set of parameters from which the original parameter cannot be recovered is a set of Lebesgue measure zero in the parameter space. In our case, although we cannot guarantee that $\mathcal{C}(S_1) \cap \mathcal{C}(S_2) = \emptyset$, we will show that if we select parameters for either model, the resulting distribution will lie in $\mathcal{C}(S_1) \cap \mathcal{C}(S_2)$ with probability zero.

In Chifman and Kubatko (2015), it was shown that the unrooted topology of the tree parameter for the coalescent model is generically identifiable when the nucleotide substitution model is GTR+I+ Γ or any of the commonly used submodels thereof, using the machinery of analytic functions and varieties. A function f with domain an open set $U \subseteq \mathbb{R}^m$ and range \mathbb{R} is *real analytic* on U if it is given locally by a convergent power series. An analytic variety is the common zero set of a collection of analytic functions. For the purposes of this paper, we will only need to consider analytic varieties defined by a single function, that is, varieties of the form

$$\mathcal{V}(f) = \{\mathbf{u} \in U \mid f(\mathbf{u}) = 0\},$$

where f is real analytic on U . The property of real analytic functions that we will use later is the following: For a real analytic function f with domain an open set $U \subseteq \mathbb{R}^m$, either f is identically zero or $\mathcal{V}(f)$ is a set of Lebesgue measure zero (Mityagin 2015).

To illustrate how we will use this property, we describe the strategy used in Chifman and Kubatko (2015) to prove the generic identifiability of the unrooted topology of the species tree parameter of the coalescent model. For the coalescent model with underlying κ -state nucleotide substitution model on an n -leaf rooted species tree S , let

$$\psi_S : \Theta_S \mapsto \Delta^{\kappa^n - 1}$$

be the map from the continuous parameter space for S , Θ_S , to the probability simplex with $\text{Im}(\psi_S) = \mathcal{C}(S)$. Label the states of the model by the natural numbers $\{1, \dots, \kappa\}$. Given any two rooted species trees S_1 and S_2 that are topologically distinct when the root vertex of each is suppressed, the strategy is to find a polynomial

$$g \in \mathbb{R}[q_{i_1 \dots i_n} : 1 \leq i_1, \dots, i_n \leq \kappa]$$

such that for all $p_1 \in \mathcal{C}(S_1)$, $g(p_1) = 0$, but for which there exists $p_2 \in \mathcal{C}(S_2)$ such that $g(p_2) \neq 0$. Then, since $g(p_1) = 0$ for all $p_1 \in \mathcal{C}(S_1)$, the set of parameters in Θ_{S_2} mapping into $\mathcal{C}(S_1) \cap \mathcal{C}(S_2)$, must be contained in the zero set of

$$g \circ \psi_{S_2} : \Theta_{S_2} \rightarrow \mathbb{R}.$$

If it can then be shown that $g \circ \psi_{S_2}$ is a real analytic function, then its zero set is the analytic variety $\mathcal{V}(g \circ \psi_{S_2})$. The existence of p_2 implies that $g \circ \psi_{S_2}$ is not identically zero on Θ_{S_2} , and so the set of parameters in Θ_{S_2} mapping into $\mathcal{C}(S_1) \cap \mathcal{C}(S_2)$ must be measure zero. Doing this for all pairs of n -leaf trees that are topologically distinct when the root vertex of each is suppressed establishes the generic identifiability of the unrooted topology of the species tree parameter of \mathcal{C}_n .

We will show that the species tree parameter of each of the modified models introduced above is generically identifiable using the same approach. In the discussion proceeding (Chifman and Kubatko 2015, Corollary 1), it was shown that for the equidistant multispecies coalescent, to establish identifiability of the species tree parameter of the coalescent model for trees with any number of leaves, it is enough to prove the identifiability of the species tree parameter for the four-leaf model. Essentially, the same proof of that theorem applies to the clockless coalescent giving us the following proposition.

Proposition 3.1 *If the unrooted topology of the species tree parameter of \mathcal{C}_4^* is generically identifiable, then the unrooted topology of the species tree parameter of \mathcal{C}_n^* is generically identifiable for all n .*

A similar proposition holds for the p -coalescent but a slight modification is required. The subtlety is illustrated in Fig. 4 where a species tree and its restriction to a four-leaf subset of the leaves are shown. Notice that on the restricted tree, the effective population size may now vary within a single branch. Therefore, to show the identifiability of the unrooted species tree parameter of the p -coalescent for n -leaf trees, we must show the identifiability of the unrooted topology of the species tree parameter of a model on

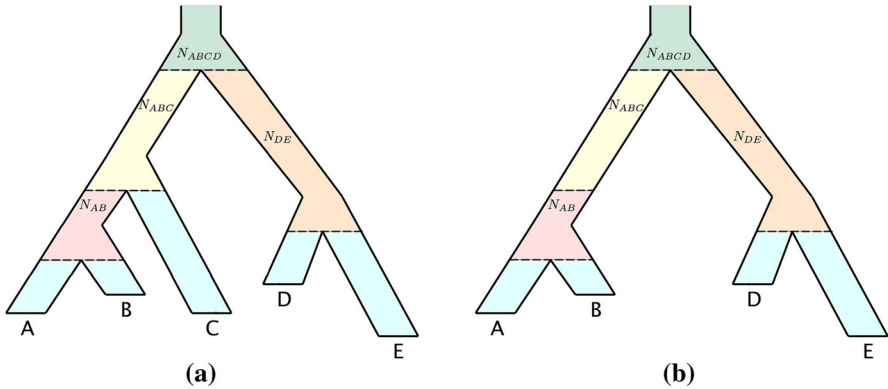


Fig. 4 A five-leaf species tree with topology S with multiple effective population size parameters and its restriction to the four-leaf topological subtree $S|_{\{A, B, D, E\}}$. The image of the marginalization map applied to the model for S will be the model for $S|_{\{A, B, D, E\}}$ with different effective population size parameters on different portions of e_{AB}

four-leaf trees that allows for a finite number of bands on each branch with separate effective population sizes. We will revisit this point after the proof of Theorem 3.5, though it turns out to be rather inconsequential.

3.1 The analyticity of ψ_S

In the discussion preceding Proposition 3.1, we described how to use the properties of real analytic varieties to prove generic identifiability. One of the results needed was that the function $g \circ \psi_S$ is a real analytic function. Since polynomial functions are real analytic and the composition of real analytic functions is again analytic, to prove this it is enough to show that for any tree S , each coordinate of ψ_S is a real analytic function in the continuous parameters of the model. That this is so may seem obvious to some and was stated without proof in Chifman and Kubatko (2015). However, this issue is slightly more subtle than it might first appear.

Recall that each coordinate of $\psi_S : \Theta_S \rightarrow \Delta^{\kappa^n - 1}$ is defined by a function of the form

$$\sum_{h \in \mathcal{H}} \int_{\mathbf{t}} p_{i_1 \dots i_n}^* |(G, \mathbf{t})| f_h((G, \mathbf{t}) | (S, \boldsymbol{\tau})) d\mathbf{t}.$$

The entries of the matrix exponential are defined by convergent power series on Θ_S and so are real analytic functions on Θ_S . Moreover, since elementary functions are analytic, as are sums, products, and compositions of real analytic functions Krantz and Parks (2002), the function $p_{i_1 \dots i_n}^* |(G, \mathbf{t})| f_h((G, \mathbf{t}) | (S, \boldsymbol{\tau}))$ is also a real analytic function on the entire parameter space. However, notice that the integral may be improper as in Example 2.1. It is not in general true that taking an improper integral with respect to certain variables in a real analytic function results in a real analytic function. As

a counterexample, consider the function $f(\alpha, t) = \frac{d}{dt}(\alpha \tanh(\alpha t))$ and define the function $F(\alpha) = \int_0^\infty f(\alpha, t) dt$. Then, $f(\alpha, t)$ is a real analytic function on its entire domain, but $F(\alpha) = |\alpha|$ and so is not analytic at $\alpha = 0$.

For the models JC69, K2P, K3P, F81, HKY85, TN93, and the generalized κ -state JC, these issues become irrelevant, as we can diagonalize the rate matrices and obtain a closed-form expression for the entries of the transition matrices. The entries are then seen to be exponential functions of branch length, and we can solve the improper integrals from the multispecies coalescent and obtain exact formulas for each coordinate of ψ_S that are clearly analytic. Thus, we have the following proposition.

Proposition 3.2 *Let S be a rooted four-leaf species tree. The parameterization map ψ_S is analytic when the underlying nucleotide substitution model is any of JC69, K2P, K3P, F81, HKY85, TN93, or the generalized κ -state JC.*

The rate matrix for the four-state general time-reversible model is similar to a real symmetric matrix and is thus also diagonalizable. However, actually writing down a closed form for the entries of the transition matrix is not possible due to the large number of computations involved. Consequently, we cannot write down a closed-form expression for the coordinate functions of ψ_S . Of course, this is not a necessary condition for these functions to be analytic, but it is difficult to argue that they are without such a closed-form expression. Therefore, in the proposition below, we will argue that around a generic choice of parameters for the GTR rate matrix, there exists a neighborhood on which the entries of the matrix exponential can be written as expressions involving only elementary functions of the rate matrix parameters, roots of the rate matrix parameters, and exponential functions. This allows us to argue that the coordinate functions of ψ_S can also be expressed in terms of well-known functions of the rate matrix parameters, and hence, that they are real analytic functions in a neighborhood around any generic choice of parameters from the modified coalescent models.

Proposition 3.3 *Let S be a rooted four-leaf species tree. Let ψ_S be the parameterization map for the multispecies coalescent model when the underlying nucleotide substitution model is the four-state GTR model. For a generic choice of continuous parameters $\theta \in \Theta_S$, there exists a neighborhood around θ on which each coordinate of ψ_S is a real analytic function.*

Proof Let θ be a generic point in Θ_S , and let Q be the rate matrix for the four-state GTR model. The matrix

$$A = \text{diag}(\pi^{1/2})Q\text{diag}(\pi^{-1/2})$$

is a real symmetric matrix that is similar to Q . Hence, all eigenvalues of A are real numbers that are less than or equal to zero and one of these eigenvalues is $\lambda_1 = 0$. We can factor the degree four characteristic equation of A and use the cubic formula to write the other eigenvalues $0 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4$ in terms of the rate matrix parameters. For a generic choice of parameters, the eigenvalues will be distinct and the columns

of $\prod_{j \neq i} (A - \lambda_j I)$ will be eigenvectors of A with eigenvalue λ_i . Define the vector V_i to be the first column of $\prod_{j \neq i} (A - \lambda_j I)$ for $1 \leq i \leq 4$, and let U be the 4×4 matrix with i -th column equal to $V_i / \|V_i\|$. Since A is a real symmetric matrix, the eigenvectors corresponding to distinct eigenvalues are orthogonal Hoffman and Kunze (1971); hence, U is an orthonormal matrix. Therefore, $A = U \text{diag}(0, \lambda_2, \lambda_3, \lambda_4) U^T$ and

$$e^{Qt} = (\text{diag}(\boldsymbol{\pi}^{-1/2})U)\text{diag}(1, e^{\lambda_2 t}, e^{\lambda_3 t}, e^{\lambda_4 t})(U^T \text{diag}(\boldsymbol{\pi}^{1/2})).$$

Thus, in a neighborhood around $\boldsymbol{\theta}$, each entry of the matrix exponential can be written as

$$P_{ij}(t) = \sum_{1 \leq k \leq 4} f_k^{(ij)}(q) e^{\lambda_k t},$$

where the $f_k^{(ij)}(q)$ are rational functions of the rate matrix parameters and roots of the rate matrix parameters coming from the cubic formula.

The functions $p_{i_1 \dots i_n}^*(G, \mathbf{t})$ are all sums of products of these functions which are exponential in the branch length t . The formulas coming from the coalescent process, $f_h((G, \mathbf{t})|(S, \boldsymbol{\tau}))$, are also exponential functions in t . Because each λ_i is guaranteed to be less than or equal to zero, when we integrate each $p_{i_1 \dots i_n}^*(G, \mathbf{t}) f_h((G, \mathbf{t})|(S, \boldsymbol{\tau}))$ with respect to branch length, the integral converges. Therefore, in a neighborhood around $\boldsymbol{\theta}$, each coordinate of ψ_S can be written in closed form as an expression involving rational functions of the model parameters, roots of the model parameters, and exponential functions of both of these. □

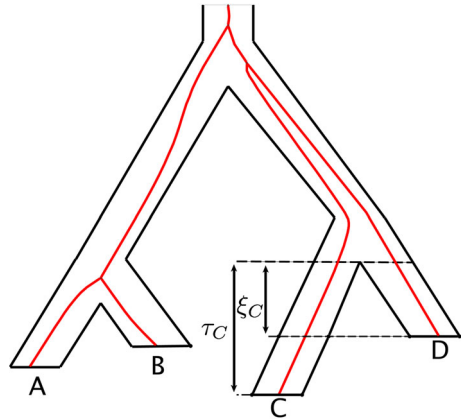
3.2 Identifiability of the Modified Multispecies Coalescent for Four-Leaf Trees

We may encode the site pattern probability distribution associated with a κ -state phylogenetic model on an n -leaf species tree as an n -dimensional $\kappa \times \dots \times \kappa$ tensor P where the entry $P_{i_1 \dots i_n}$ is the probability of observing the state $i_1 \dots i_n$. In (Chifman and Kubatko 2015, Section 4), the authors explain how to construct tensor flattenings according to a bipartition of the taxa, or split, of the species tree. Our first result is the analogue of (Chifman and Kubatko 2015, Theorem 1) for the modified coalescent models. We use the notation $P_{(S, \boldsymbol{\tau}, \boldsymbol{\theta})}$ to denote the probability tensor that results from choosing a species tree S with vector of edge lengths $\boldsymbol{\tau}$ and continuous parameters $\boldsymbol{\theta}$.

Theorem 3.4 *Let S be a four-taxon symmetric $((A, B), (C, D))$ or asymmetric $(A, (B, (C, D)))$ species tree with a cherry (C, D) . Consider the clockless coalescent when the underlying nucleotide substitution model is any of the following: JC69, K2P, K3P, F81, HKY85, TN93, or GTR. Let $L_1|L_2$ be the split $AB|CD$ that is valid for S . Then, for all $P_{(S, \boldsymbol{\tau}, \boldsymbol{\theta})} \in \mathcal{C}^*(S)$,*

$$\text{rank}(\text{Flat}_{L_1|L_2}(P_{(S, \boldsymbol{\tau}, \boldsymbol{\theta})})) \leq 10.$$

Fig. 5 Extending one leaf in a cherry of S



Proof Let $L_1|L_2$ be the split $AB|CD$ that is valid for S , and consider the distribution $P_{(S, \tau, \theta)}$. Without loss of generality, suppose $\tau_C \geq \tau_D$. Consider the new vector of edge lengths ξ where each entry is the same as in τ except that $\xi_C = \tau_D$. Thus, we can think of the tree (S, τ) as an extension of the tree (S, ξ) as in Fig. 5.

First, we claim that

$$\text{rank}(\text{Flat}_{L_1|L_2}(P_{(S, \tau, \theta)})) \leq \text{rank}(\text{Flat}_{L_1|L_2}(P_{(S, \xi, \theta)}))$$

Notice that since coalescent events do not happen in the terminal edges of the species tree, the gene tree histories and the formulas for the gene tree distributions from (S, τ) and (S, ξ) are identical. The only difference is that the leaf edge labeled by C in the gene tree (G, \mathbf{t}) from (S, τ) is longer by $\tau_C - \xi_C$ than the same edge in the gene tree (G, \mathbf{t}) from (S, ξ) . The probability of observing the state $i_1i_2i_3i_4$ from these two gene trees will not be the same, so let us express this probability as $p_{i_1i_2i_3i_4|G, \mathbf{t}, \theta}^*$ when the species tree is (S, ξ) and $q_{i_1i_2i_3i_4|G, \mathbf{t}, \theta}^*$ when the species tree is (S, τ) . Extending the branch of a gene tree is equivalent to grafting a new edge onto the leaf edge to create an internal vertex of degree two. To compute the probability of observing a particular state at the leaves of the extended gene tree, we sum over all possible states of this vertex. For clarity of notation, let us represent the matrix of transition probabilities along the grafted edge by $M = e^{Q(\tau_C - \xi_C)}$. Thus,

$$q_{i_1i_2i_3i_4|G, \mathbf{t}, \theta}^* = \sum_{1 \leq j \leq 4} (M_{ji_3}) p_{i_1i_2ji_4|G, \mathbf{t}, \theta}^*$$

Therefore, the total probability for a particular history is given by

$$\begin{aligned} p_{i_1i_2i_3i_4|h, (S, \tau, \theta)} &= \int_{\mathbf{t}} q_{i_1i_2i_3i_4|G, \mathbf{t}, \theta}^* f_h((G, \mathbf{t})|(S, \tau)) \, d\mathbf{t} \\ &= \int_{\mathbf{t}} \left(\sum_{1 \leq j \leq 4} (M_{ji_3}) p_{i_1i_2ji_4|G, \mathbf{t}, \theta}^* \right) f_h((G, \mathbf{t})|(S, \xi)) \, d\mathbf{t} \end{aligned}$$

$$\begin{aligned}
 &= \sum_{1 \leq j \leq 4} (M_{j i_3}) \left(\int_{\mathbf{t}} p_{i_1 i_2 j i_4 | (G, \mathbf{t}, \theta)}^* f_h((G, \mathbf{t}) | (S, \xi)) \, d\mathbf{t} \right) \\
 &= \sum_{1 \leq j \leq 4} (M_{j i_3}) p_{i_1 i_2 j i_4 | h, (S, \xi, \theta)}
 \end{aligned}$$

Summing over all histories, we also obtain

$$p_{i_1 i_2 i_3 i_4 | (S, \tau, \theta)} = \sum_{1 \leq j \leq 4} (M_{j i_3}) p_{i_1 i_2 j i_4 | (S, \xi, \theta)}.$$

Now consider the column of $Flat_{L_1 | L_2}(P_{(S, \tau, \theta)})$ indexed by the joint state $i_3 i_4$. The formula above shows that this column is a linear combination of the columns of $Flat_{L_1 | L_2}(P_{(S, \xi, \theta)})$ indexed by $1i_4, 2i_4, 3i_4,$ and $4i_4$. Therefore,

$$rank(Flat_{L_1 | L_2}(P_{(S, \tau, \theta)})) \leq rank(Flat_{L_1 | L_2}(P_{(S, \xi, \theta)})).$$

Thus, any four-leaf species tree (S, τ) with a (C, D) cherry can be constructed by lengthening one terminal edge in a tree (S, ξ) with a (C, D) cherry that satisfies $\xi_C = \xi_D$. The tree (S, ξ) may not be equidistant, but it is still clear from the symmetry in the cherry that for any choice of continuous parameters, we will have

$$p_{i_1 i_2 i_3 i_4 | (S, \xi, \theta)} = p_{i_1 i_2 i_4 i_3 | (S, \xi, \theta)},$$

which implies that $rank(Flat_{L_1 | L_2}(P_{(S, \xi, \theta)})) \leq 10$, and hence that $rank(Flat_{L_1 | L_2}(P_{(S, \tau, \theta)})) \leq 10$. □

Theorem 3.5 *Let S be a four-taxon symmetric $((A, B), (C, D))$ or asymmetric $(A, (B, (C, D)))$ species tree with a cherry (C, D) . Let $L_1 | L_2$ be one of the splits $AC | BD$, or $AD | BC$. Consider the clockless coalescent when the underlying nucleotide substitution model is any of the following: JC69, K2P, K3P, F81, HKY85, TN93, or GTR. Then, for generic distributions $P_{(S, \tau, \theta)} \in \mathcal{C}^*(S)$,*

$$rank(Flat_{L_1 | L_2}(P_{(S, \tau, \theta)})) = 16.$$

Proof Consider the degree 16 polynomial $\det(Flat_{L_1 | L_2}(q))$ in the ring $\mathbb{R}[q_{i_1 \dots i_n} : 1 \leq i_1, \dots, i_4 \leq 4]$. Any choice of continuous parameters for the clockless coalescent that satisfies

$$rank(Flat_{L_1 | L_2}(P_{(S, \tau, \theta)})) < 16,$$

must be contained in the real analytic variety $\mathcal{V}(\det(Flat_{L_1 | L_2}(q)) \circ \psi_S)$. As per the discussion at the beginning of Section 3, to show that this is a set of measure zero, we need only verify that the function $\det(Flat_{L_1 | L_2}(q)) \circ \psi_S$ is not identically zero. To do so, we need only produce a single choice of parameters for both the symmetric and asymmetric trees for which $Flat_{AC | BD}(P_{(S, \tau, \theta)})$ is rank 16 and likewise for

$Flat_{AD|BC}(P_{(S,\tau,\theta)})$. In fact, we can address both the symmetric and asymmetric cases with one tree by letting S be the symmetric tree and setting $\tau_{AB} = 0$. In the supplemental materials, we choose parameters from the Jukes–Cantor model and show that both flattening matrices for the invalid splits are rank 16. Since the Jukes–Cantor model is contained in JC69, K2P, K3P, F81, HKY85, and TN93, this choice of parameters establishes the result for each of these.

The Jukes–Cantor model is of course also contained in the four-state GTR model. However, in light of Proposition 3.3, our choice of parameters must be sufficiently generic so that $\det(Flat_{L_1|L_2}(q) \circ \psi_S)$ is a real analytic function in a neighborhood around this point. In the supplemental materials, we also choose a set of sufficiently generic K3P parameters and show that both flattening matrices for the invalid splits are rank 16. \square

Suppose now that S is a four-leaf tree that displays the split $L_1|L_2$ and that S' is a four-leaf tree with different unrooted topology than S . Then, S' does not display the split $L_1|L_2$, and Theorem 3.5 shows that the determinant of $Flat_{L_1|L_2}(q)$ is a degree 16 polynomial that does not vanish on the set $\mathcal{C}^*(S')$. Hence, as per the discussion in Sect. 3, the set of parameters for S' mapping into $\mathcal{C}^*(S) \cap \mathcal{C}^*(S')$ is a set of measure zero. Thus, the unrooted topology of the species tree parameter of the clockless coalescent is generically identifiable.

Following Proposition 3.1, we observed that showing the identifiability of the unrooted topology of the species tree parameter of the p -coalescent requires proving the identifiability of the unrooted topology of the species tree parameter for four-leaf trees in a model that allows multiple effective population size parameters on a single edge. Specifically, to prove the identifiability of the unrooted topology of the species tree parameter in $\mathcal{C}_n^*(N)$, it is sufficient to prove the identifiability of the unrooted topology of the species tree parameter for four-leaf trees in a model with $2n - 3$ different effective population size parameters on each edge. This is because the effective population size parameters in the four-leaf tree are inherited from the original n -leaf tree, and the number of different effective population size parameters in an n -leaf tree is bounded above by the number of edges, $2n - 3$. All of the key ingredients needed to prove this result have already been presented in Theorems 3.4 and 3.5.

Since coalescent events do not occur in the terminal edges of the species tree, Theorem 3.4 applies equally to the p -coalescent and clockless p -coalescent models. Both distributions in the proof of Theorem 3.5 are still contained in the model where we allow multiple effective population size parameters on each edge since we can just choose all of the population size parameters on e_{CD} to be equal. We must still verify that the parameterization map for this model is analytic, but the argument from Sect. 3.1 remains unchanged when we allow multiple effective population size parameters on each edge. Thus, the same choices of parameters from the proof of Theorem 3.5 establish the result for the clockless p -coalescent. We also intentionally chose a point corresponding to an equidistant tree so that it applies to the p -coalescent. Thus, we have the following corollary.

Corollary 3.6 *The unrooted topology of the species tree parameter of the clockless coalescent, the p -coalescent, and the clockless p -coalescent models on an n -leaf tree*

is generically identifiable for all n when the underlying nucleotide substitution model is any of the following: JC69, K2P, K3P, F81, HKY85, TN93, or GTR.

3.3 Identifiability with Invariable Sites and Gamma-Distributed Rates

It is well known that the rate of evolution may vary across sites (Yang 1993, 1994). One way to account for this is to let each site evolve according to the same model but where the rate matrix at each site is scaled by a random factor drawn from a specified gamma distribution. If the underlying nucleotide substitution model is assumed to be the GTR model, this is what is known as the GTR+ Γ model.

In practice, the gamma distribution is approximated using m rate categories, each with probability $\frac{1}{m}$, and ρ_i is defined to be the mean rate for category i (see Yang 1994 for details). From the formulas in Yang (1994), it is easy to see that the rates can be expressed as analytic functions in the parameters of the gamma distribution and consequently that the distributions from the GTR+ Γ model are given by real analytic functions of the parameters.

It is also common to account for invariable sites by using the GTR+I+ Γ model, where δ is the proportion of invariable sites. The multispecies coalescent with the m -discrete κ -state GTR+I+ Γ model was shown to exhibit the same flattening ranks as the multispecies coalescent with the κ -state GTR model in Chifman and Kubatko (2015). This is not terribly surprising as a probability distribution from the former is the sum of $m + 1$ distributions each satisfying the same linear relations. Explicitly, letting $P^{I+\Gamma}$ be the site pattern probability distribution from a model with invariant sites and gamma-distributed rates,

$$P_{i_1 i_2 i_3 i_4}^{I+\Gamma}(S, \tau, \theta) = \frac{(1 - \delta)}{m} (p_{i_1 i_2 i_3 i_4}^{\rho_1}(S, \tau, \theta) + \dots + p_{i_1 i_2 i_3 i_4}^{\rho_m}(S, \tau, \theta)) + \delta(z_{i_1 i_2 i_3 i_4}|\theta),$$

where $p_{i_1 i_2 i_3 i_4}^{\rho_j}(S, \tau, \theta)$ is the probability of observing $i_1 i_2 i_3 i_4$ from the multispecies coalescent model with scaling factor ρ_j and $z_{i_1 i_2 i_3 i_4}|\theta$ is the probability of observing this state at an invariable site. If S has a (C, D) cherry as above, then each summand is contained in the linear space defined by the linear relations of the form $p_{**i_3 i_4} - p_{**i_4 i_3}$ in the distribution space. The sum satisfies these relations as well, so we have

$$\text{rank}(Flat_{AB|CD}(P_{(S, \tau, \theta)}^{I+\Gamma})) \leq \binom{\kappa + 1}{2}.$$

For a non-equidistant tree, the same result no longer applies. If we view (S, τ) as an extension of (S, ξ) as we did in Theorem 3.4, we can see that

$$\text{rank}(Flat_{AB|CD}(P_{(S, \tau, \theta)}^{\rho_j})) \leq \binom{\kappa + 1}{2},$$

but the particular linear relationships satisfied by the columns of each flattening matrix will depend on the entries of the transition matrix on the extended edge, which in turn

depend on the ρ_i . However, we can obtain an analogous result for JC+I+ Γ , where the JC refers to the κ -state Jukes–Cantor model. When $\kappa = 4$, we prove the result for $m = 2, 3$, and 4, as four is the most common number of categories used in actual phylogenetic applications (Lio and Goldman 1998).

Theorem 3.7 *Let S be a four-taxon symmetric $((A, B), (C, D))$ or asymmetric $(A, (B, (C, D)))$ species tree with a cherry (C, D) . Let $L_1|L_2$ be one of the splits $AB|CD, AC|BD$, or $AD|BC$. For $\kappa \geq 4$, consider the κ -state m -discrete JC+I+ Γ model under the coalescent with species tree S and $m \leq 4$.*

(1) *If $L_1|L_2$ is a valid split for S , then for all $P_{(S,\tau,\theta)}^{I+\Gamma}$ from the clockless p -coalescent with invariant sites and gamma-distributed rates,*

$$\text{rank}(\text{Flat}_{L_1|L_2}(P_{(S,\tau,\theta)}^{I+\Gamma})) \leq \kappa^2 - \binom{\kappa - 1}{2}$$

(2) *If $L_1|L_2$ is not a valid split for S , then for a generic distribution $P_{(S,\tau,\theta)}^{I+\Gamma}$ from the clockless p -coalescent with invariant sites and gamma-distributed rates,*

$$\text{rank}(\text{Flat}_{L_1|L_2}(P_{(S,\tau,\theta)}^{I+\Gamma})) > \kappa^2 - \binom{\kappa - 1}{2}.$$

Proof Let $L_1|L_2$ be the split $AB|CD$ that is valid for S , and consider the distribution $P_{(S,\tau,\theta)}$ from the Jukes-Cantor model. Without loss of generality, suppose $\tau_C \geq \tau_D$. Construct the vector ξ with all entries equal to those of τ but with $\xi_C = \tau_D$. Again, by symmetry, we have

$$P_{\star\star i_3 i_4 | (S, \xi, \theta)} = P_{\star\star i_4 i_3 | (S, \xi, \theta)}.$$

As in Theorem 3.4, we will identify the tree (S, τ) as an extension of (S, ξ) . For the JC model, there are only two distinct entries of $M = e^{Q(\tau_C - \xi_C)}$. Let $M_{ij} = a$ if $i = j$ and b otherwise. Therefore, we have

$$P_{\star\star i_3 i_4 | (S, \tau, \theta)} = a P_{\star\star i_3 i_4 | (S, \xi, \theta)} + \sum_{j \neq i_3} b P_{\star\star j i_4 | (S, \xi, \theta)},$$

and one can check that for distinct $k_1, k_2, k_3 \in [\kappa]$, the distribution $P_{|(S,\tau,\theta)}$, satisfies

$$P_{\star\star k_1 k_2} - P_{\star\star k_1 k_3} - P_{\star\star k_2 k_1} + P_{\star\star k_2 k_3} + P_{\star\star k_3 k_1} - P_{\star\star k_3 k_2} = 0.$$

We obtain such a relation for any three-element subset of $[\kappa]$. Moreover, since this linear relation does not depend on a or b , it is satisfied by $P_{|(S,\xi,\theta)}^{\rho_i}$. It is also satisfied by the matrix for invariable sites, $Z_{|\theta}$, with entries given by $z_{i_1 i_2 i_3 i_4 | \theta}$. Hence, this linear relation is also satisfied by any distribution from the m -discrete JC+I+ Γ model. Consider the $\binom{\kappa - 1}{2}$ relations that come from choosing three-element subsets of the form $\{k_1, k_2, \kappa\}$. For all $k_1, k_2 \in [\kappa - 1]$, exactly one of these relations involves the

variable $p_{**k_1k_2}$. Therefore, these relations are linearly independent, and so the first claim of the theorem follows.

In (Chifman and Kubatko 2015, Theorem 1), the authors show that for all m , when $\kappa \geq 4$, if $L_1|L_2$ is not a valid split for S , then

$$\text{rank}(\text{Flat}_{L_1|L_2}(P_{(S,\tau,\theta)}^{I+\Gamma})) > \kappa^2 - \kappa.$$

When $\kappa \geq 5$, we have

$$\kappa^2 - \kappa \geq \kappa^2 - \binom{\kappa - 1}{2},$$

which establishes our result. For $\kappa = 4$, we must produce a choice of parameters to prove that the claim holds for $m = 2, 3$, and 4 and for both the symmetric and asymmetric trees. Choosing $\alpha = \beta = 1$, $\delta = 1/2$, and the same continuous JC69 parameters from Theorem 3.5 establishes the result. Code to verify these computations is contained in the supplementary materials. \square

Since all of the parameterization functions involved are analytic, this is enough to prove the identifiability of the unrooted topology of the species tree parameter of the JC+I+ Γ model. Thus, we have the following corollary.

Corollary 3.8 *The unrooted topology of the species tree parameter of the clockless coalescent, the p -coalescent, and the clockless p -coalescent models on an n -leaf tree is generically identifiable for all n when the underlying nucleotide substitution model is the m -discrete κ -state JC+I+ Γ model with $\kappa \geq 5$ and $m \in \mathbb{N}$ and with $\kappa = 4$ and $m = 2, 3$, or 4.*

Moreover, the parameters that we used to demonstrate that the invalid flattenings are full rank come from an exponential distribution, which is a special case of the gamma distribution. Therefore, the same result holds for a model where the m rates are constructed from an exponential distribution. In fact, this also applies to a more general variable rates model where the m rates are free parameters.

4 Conclusions

In the previous section, we have proven that the unrooted species tree parameter of several more generalized versions of the multispecies coalescent model is generically identifiable from the site pattern probability distributions on the species trees. Moreover, the means by which we have proven identifiability give us the necessary framework for reconstructing the unrooted topology of the species tree from data. In each case, we showed that we can reconstruct the unrooted quartets of the species tree parameter if we know the distribution exactly by taking ranks of the flattening matrices. Specifically, for a four-state model and generic choices of parameters, we showed that the rank of the flattening matrix for the quartet compatible with the species tree

will be less than or equal to 10 while the other two flattening matrices will both be rank 16.

This gives a natural method for inferring the unrooted topology of the species from biological data. Specifically, for each quartet, we infer the unrooted quartets of the species tree by determining which of the three flattening matrices is closest to the set of rank 10 matrices. The method of singular value decomposition from linear algebra already provides a means of determining how close a matrix is to the set of matrices of a certain rank under the Frobenius norm (Golub and Loan 2013). This is exactly the procedure used by the method SVDQuartets, which is already fully implemented in the PAUP* software (Swofford 2016). Hence, there is strong theoretical justification for applying SVDQuartets for phylogenetic reconstruction even when effective population sizes vary throughout the tree or when the molecular clock does not hold.

The model presented in Sect. 2, as well as that presented in Chifman and Kubatko (2015), describes the situation in which gene trees are randomly sampled under the multispecies coalescent model, and then sequence data for a single site evolve along each sampled gene tree according to one of the standard nucleotide substitution models. Data generated in this way have been termed “coalescent independent sites” Tian and Kubatko (2016) to distinguish them from SNP data. Although coalescent independent sites and SNP data refer to observations of single sites that are assumed to be conditionally independent samples from the model given the species tree, SNP data are generally biallelic, while coalescent independent sites may include three or four nucleotides at a site, or may be constant.

The other situation in which one might wish to apply these results is to multilocus data. Multilocus data are data in which individual genes are sampled from the species tree under the multispecies coalescent, but for each sampled gene tree, many individual sites are observed. Typical genes observed in phylogenomic studies range from 100 base pairs (bp) to 2000 bp in size, though most are < 500 bp. The site patterns observed within a gene are not independent observations under the model because they share the same gene tree, and thus it is not immediately obvious that the results presented here apply to this case. However, consider the case in which a large sample of genes, say W , is obtained, and for each gene, s sites are observed. Then, the flattening matrices of site pattern counts constructed from such data will be s times the flattening matrix of site pattern counts that would have been observed if only a single site had been observed from each gene tree, which does not change the matrix rank. It is clear that as $W \rightarrow \infty$, the correct theoretical distribution will be well approximated by the observed site pattern frequencies, and the results presented here will hold. In practice, the genes will vary in their lengths and a more careful argument is required. We have elsewhere carried out thorough simulation studies to show that the methods used in SVDQuartets hold for multilocus data as well as for SNP data and for coalescent independent sites for the original model (Chifman and Kubatko 2014). We are currently working on a simulation study to compare the effectiveness of SVDQuartets to that of other species tree estimation methods on the models presented herein.

We note two possible criticisms of this method. The first is that, while we showed that generically the flattening matrices for the invalid splits will be rank 16, we have no theoretical guarantees that they are not arbitrarily close to the set of rank 10 matrices. Therefore, we do not know a priori that this method will provide any insight with a

finite amount of either simulated or biological data. Along the same lines, determining that a flattening matrix is close to the set of rank 10 matrices does not necessarily mean that it is close to the set of distributions arising from a coalescent model, as the latter is properly contained in the former. While both are valid considerations, they appear to be academic, as SVDQuartets has already been shown to be an effective reconstruction method on several data sets, both real and simulated (Chifman and Kubatko 2014; Chou et al. 2015). As mentioned above, in a forthcoming paper, we will demonstrate that SVDQuartets also works well in practice by simulating data from these modified coalescent models and applying the method to real biological data sets known to violate the molecular clock.

In recent years, the amount of sequence data available for species tree inference has increased rapidly, presenting significant computational challenges for most model-based species tree inference methods that accommodate the coalescent process. The SVDQuartets method is fully model based but inference using this method is much more computationally efficient than methods that require evaluation of a likelihood function, such as *BEAST Heled and Drummond (2010) and SNAPP Bryant et al. (2012). This is because, for each quartet considered, all that is required is construction of the three flattening matrices, which involves the simple task of counting site patterns and computation of singular values from these 16×16 matrices. In addition, increases in sequence length benefit the performance of the method (because site pattern probabilities are estimated more accurately) with almost no increased computational cost. However, increases in the number of sequences do incur a computational cost, in that more quartets must be evaluated and because the complexity of the algorithm for assembling the inferred quartets to form an overall species tree estimate increases. Even with these costs, however, computations can be carried out much more rapidly than with the likelihood-based methods referenced above. In the work presented here, we show that the theory underlying the SVDQuartets method holds in much more general settings than originally suggested. In particular, the method can be applied to data that violate the molecular clock and to the case in which each population has a distinct effective population size. Thus, this work is a significant advance that will contribute meaningfully to the collection of methods available to infer species-level phylogenies from phylogenomic data in very general settings.

Acknowledgements This research has been supported in part by the Mathematical Biosciences Institute and the National Science Foundation under Grant DMS 1440386. We would like to thank the two anonymous reviewers for several helpful remarks.

References

- Bryant D, Bouckaert R, Felsenstein J, Rosenberg N, Roy Choudhury A (2012) Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol Biol Evol* 29(8):1917–1932
- Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10:195–205
- Chifman J, Kubatko L (2014) Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30(23):3317–3324

- Chifman J, Kubatko L (2015) Identifiability of the unrooted species tree topology under the coalescent model with time specific rate variation and invariable sites. *J Theor Biol* 374:35–47
- Chou J, Gupta A, Yaduvanshi S, Davidson R, Nute M, Mirarab S, Warnow T (2015) A comparative study of SVDQuartets and other coalescent-based species tree estimation methods. *BMC Genom* 16(Suppl 10):S2
- Degnan J, Salter L (2005) Gene tree distributions under the coalescent process. *Evolution* 59:24–37
- Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17(6):368–76
- Golub GH, Loan CFV (2013) Matrix computation. Johns Hopkins University Press, 4th edn. Section 2.4
- Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. *Mol Biol Evol* 27(3):570–580
- Hoffman K, Kunze R (1971) Linear algebra, 2nd edn. Prentice Hall, New Jersey
- Kingman JFC (1982) Exchangeability and the evolution of large populations. In: Koch G, Spizzichino F (eds) Exchangeability in probability and statistics. North-Holland, Amsterdam, pp 97–112
- Kingman JFC (1982) On the genealogy of large populations. *J Appl Prob* 19A:27–43
- Kingman JFC (1982) The coalescent. *Stoch Proc Appl* 13:235–248
- Krantz SG, Parks HR (2002) A primer of real analytic functions, 2nd edn. Springer, New York
- Lio P, Goldman N (1998) Models of molecular evolution and phylogeny. *Genome Res* 8:1233–1244
- Liu L, Yu L, Edwards S (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol Biol* 10(1):302
- Liu L, Yu L, Pearl D, Edwards S (2009) Estimating species phylogenies using coalescence times among sequences. *Syst Biol* 58(5):468–477
- Maddison WP (1997) Gene trees in species trees. *Syst Biol* 46:523–536
- Martin AP, Palumbi SR (1993) Body size, metabolic rate, generation time, and the molecular clock. *Proc Natl Acad Sci USA* 90:4087–4091
- Mirarab S, Reaz R, Bayzid MD, Zimmermann T, Swenson MS, Warnow T (2014) Astral: genome-scale coalescent-based species tree. *Bioinformatics (ECCB special issue)* 30(17):i541–i548
- Mirarab S, Warnow T (2015) Astral-ii: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics (ISMB special issue)* 31(12):i44–i52
- Mityagin B (2015) The zero set of a real analytic function. [arXiv:1512.07276](https://arxiv.org/abs/1512.07276)
- Pamilo P, Nei M (1988) Relationships between gene trees and species trees. *Mol Biol Evol* 5(5):568–583
- Semple C, Steel M (2003) Phylogenetics. Oxford University Press, Oxford
- Swofford D (2002) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts
- Swofford D (2016) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4a150
- Syvanen M (1994) Horizontal gene transfer: evidence and possible consequences. *Annu Rev Genet* 28:237–261
- Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460
- Takahata N, Nei M (1985) Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* 110:325–344
- Tavaré S (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor Popul Biol* 26:119–164
- Tian Y, Kubatko L (2016) Rooting phylogenetic trees under the coalescent model using site pattern probabilities. (submitted)
- Wu Y (2012) Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution* 66(3):763–775
- Yang Z (1993) Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol* 10:1396–1401
- Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* 39(3):306–314